

# Capfit<sup>gen3</sup>



**A toolbox for the conservation  
and promotion of the use  
of agricultural biodiversity**



Funded by the  
Horizon 2020  
Framework  
Programme of the  
European Union



UNIVERSITY OF  
BIRMINGHAM



UNIVERSIDAD  
NACIONAL  
DE COLOMBIA



# Capfit<sup>gen3</sup>



**A toolbox for the conservation  
and promotion of the use  
of agricultural biodiversity**

## **Author of the tools**

Mauricio Parra Quijano

## **Authors of the document**

Mauricio Parra Quijano, José María Iriondo,  
María Elena Torres, Francisco López,  
Jade Phillips and Shelagh Kell



Funded by the  
Horizon 2020  
Framework  
Programme of the  
European Union



UNIVERSITY OF  
BIRMINGHAM



UNIVERSIDAD  
NACIONAL  
DE COLOMBIA

## Catalogación en la publicación Universidad Nacional de Colombia

Parra Quijano, Mauricio, 1973-

CAPFITOGEN 3 : a toolbox for the conservation and promotion of the use of agricultural biodiversity / author of the tools, Mauricio Parra Quijano ; authors of the document, Mauricio Parra Quijano [y otros cinco] ; [figures and photographs by Mauricio Parra Quijano] ; [Ana María Díaz, translation and editing]. -- First edition. --

Bogotá : Universidad Nacional de Colombia. Facultad de Ciencias Agrarias, 2022.

1 recurso en línea (303 páginas) : ilustraciones (principalmente a color), diagramas, figuras, fotografía, mapas

Incluye referencias bibliográficas al final de cada capítulo

**ISBN 978-958-505-038-9 (en línea)**

1. Conservación de los recursos genéticos 2. Agricultura de conservación 3. Germoplasma 4. Agrobiodiversidad 5. Desarrollo de programas informáticos 6. Análisis espacial I. Iriondo, José María, 1963- II. Torres, María Elena, 1972- III. Francisco López, 1976- IV. Phillips, Jade, 1990- V. Kell, Shelagh, 1962- VI. Díaz, Ana María, traductor, editor VII. Título

CDD-23 333.953416 / 2022

## CAPFITOGEN3: A toolbox for the conservation and promotion of the use of agricultural biodiversity

© Universidad Nacional de Colombia

Bogotá campus - Faculty of Agricultural Sciences

© Several authors

2021

ISBN: 978-958-505-038-9

**Dolly Montoya**

Rector

*Universidad Nacional de Colombia*

**Author of the tools:**

**Mauricio Parra Quijano**

Associate Professor

Faculty of Agricultural Sciences

Universidad Nacional de Colombia, Bogotá

**Authors of the document:**

Mauricio Parra Quijano

José María Iriondo

María Elena Torres

Francisco López

Jade Phillips

Shelagh Kell

**Design and layout by:**

Laura Londoño M.

**Figures and photographs by:**

Mauricio Parra Quijano

**No part of this manual may be reproduced in any manner without the express written permission of the copyright holder.**

**All rights reserved.**

**Bogotá, D. C., Colombia**

# Contents

	Pag.
<b>Acknowledgments</b>	<b>7</b>
<b>1. Background and tool design</b>	<b>11</b>
<b>2. CAPFITOGEN Tools v3: Features and functioning</b>	<b>19</b>
<b>3. TesTable Tool</b>	<b>39</b>
<b>4. GEOQUAL Tool</b>	<b>49</b>
<b>5. SelecVar Tool</b>	<b>61</b>
<b>6. ELCmapas Tool</b>	<b>75</b>
<b>7. ECOGEO Tool</b>	<b>91</b>
<b>8. Representa Tool</b>	<b>107</b>
<b>9. DIVmapas Tool</b>	<b>121</b>
<b>10. ColNucleo Tool</b>	<b>145</b>
<b>11. FIGS_R Tool</b>	<b>157</b>
<b>12. rLayer Tool</b>	<b>173</b>
<b>13. Complementa Tool</b>	<b>179</b>
<b>14. Bfuture Tool</b>	<b>199</b>
<b>15. Modela Tool</b>	<b>211</b>
<b>16. Mcompare Tool</b>	<b>255</b>
<b>17. Tzones Tool</b>	<b>265</b>
<b>18. Frequent errors</b>	<b>281</b>
<b>19. Annexes</b>	<b>289</b>





# Acknowledgments

## Acknowledgments

CAPFITOGEN3 development was supported by “Networking, Partnerships and Tools to Enhance *in situ* Conservation of European Plant Genetic Resources” (short name Farmer’s Pride), funded by the Horizon 2020 Framework Programme of the European Union.

CAPFITOGEN tools and their evolution, CAPFITOGEN3, are the result of continuous work since 2012 when the first two tools were conceived and designed. These tools did not come out overnight but have been under a constant process of development since 2005 when the first ELC map was obtained. Then, other useful ecogeographic applications were developed for the conservation and use of plant genetic resources for food and agriculture (PGRFA). Since 2012, there have been great achievements, but also several mistakes have been made; we have encountered some obstacles and difficulties, but we have also come across wonderful people who have contributed to make CAPFITOGEN a dream come true. I talk about a dream because these tools were literally that, a dream I had when I finished my PhD thesis. At that time, I thought that some of these methodological advances should be available to everyone and not only to a small group of future researchers who would cite my papers. Based on that dream, I assumed the premise that the effort of working on the scientific field was only compensated when progress reached people to help them improve or make their lives easier. CAPFITOGEN has been able to reach a high number of technicians and researchers who consistently conserve and use agrobiodiversity. The program has been successful at supporting all these people by allowing them to perform analyses and tasks that would not have been possible before.

First, I would like to thank Fernando Latorre (Spain), who was once Spain’s representative to the International Treaty on Plant Genetic Resources for Food and Agriculture, for his firm support of the CAPFITOGEN program and the development of the tools. Thanks are also due for the support and comments supplied by other CRF-INRA researchers, in particular, Lucía de la Rosa and Rosa García.

Thanks to the ‘fans’ of CAPFITOGEN. This group is made up of researchers and people who work for the conservation of agrobiodiversity. They have not only supported the idea of the program but also used some of its applications



at some point. Thanks for contributing to the dissemination of its potential among the PGRFA community and for looking for opportunities for greater diffusion and funding. To the people who have accompanied me in this process as co-authors of this book. The tools we introduce here are the result of your hard work and support. Txema, Elena, Francisco, Shelagh, and Nigel are the key pieces without which this puzzle would have been impossible to complete. To other researchers and recurring users who have also become fans of CAPFITOGEN such as César Tapia, Hannes Gaisberger, Joana Brehm, Lorena Marinoni, Sandibel Vera, Rosalinda González, Jade Phillips, Heli Fitzgerald, Riina Jalonen and many more.

Many thanks to everyone who has contributed to the technical part, particularly those who have developed the R packages that have been integrated into the tools. I would like to highlight the work of Robert J. Hijmans and other generous developers of R packages for spatial analysis. Robert has also permitted the distribution of WorldClim data to CAPFITOGEN users. To Marteen van Zonneveld, Evert Thomas, Stephanie Greene, Michael Mackay, and again Robert Hijmans, whom I admire for their ability to generate new methods and applications. Your ideas have inspired me to create new tools.

Finally, I would like to thank the more than 300 students, technical curators, and directors of multiple education and research centres who work on the conservation of plant genetic resources for believing in CAPFITOGEN, adopting our technology, and attending the workshops. Their experiences as users have provided us with valuable feedback and suggestions for the continuous improvement of the tools.

I would like to thank Nigel Maxted for his guidance and advice. Thanks to Ana María Díaz for their support in the translation and editing of the CAPFITOGEN3 user manual (English version), and Laura Londoño for the beautiful layout and design.



**Regional Workshop, Binzhou (China), October 2018.**



# 1 | Background and tool design

## 1.1 What are CAPFITOGEN3 tools?

The future of agriculture rests largely on the efficient conservation and sustainable use of plant genetic resources (PGR), which requires the development of cost-effective strategies to face the environmental impacts of climate change among other challenges. The CAPFITOGEN toolbox was developed to provide support to the global PGR conservation and sustainable use community by providing software tools designed to perform spatial and ecogeographic diversity analyses to facilitate more efficient and effective PGR conservation and sustainable use planning. The CAPFITOGEN tools have been providing support to PGR technicians and scientists around the world for the past nine years. Sixteen training workshops have been delivered for more than 400 trainees in four continents, and the use of the tools has been cited at least 71 times in the period 2014 to 2021. CAPFITOGEN3 is the new iteration of the CAPFITOGEN toolbox and is composed of 15 tools usable either directly on a server via an online portal, or downloadable and used in local mode on a computer hard drive.

## 1.2. History of the development of the CAPFITOGEN tools

Under the auspices of the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) and the Spanish Agency for International Cooperation for Development (AECID), two workshops were held from 2008 to 2010 on the implementation of the ITPGRFA for countries from the Group of Latin American and Caribbean Countries (GRULAC) in Cartagena de Indias (Colombia, July–August 2008) and Antigua (Guatemala, August 2010). These activities indicated the importance of contributing to the implementation of the ITPGRFA objectives within the GRULAC community through training workshops and knowledge transfer. The success of these events also reflects the close relationship and cooperation between the national programs of the GRULAC countries and Spain.

Taking this precedent and the region's necessities as a point of departure, the Program to Strengthen Capabilities in National Plant Genetic Resources Programs in Latin America was implemented. This program was originally focused on the development of appropriate technologies for countries that are extremely agrobiodiverse but have limited human and economic resources. Its objectives were to develop and transfer technology, and to provide the appropriate training to technical staff from the Latin American countries that are signatories to the Treaty. The technology developed was initially designed to support and facilitate work related to the *ex situ* conservation of plant genetic resources (PGR), and not to create new obligations or additional duties for the staff in charge of collecting, conserving, characterizing, evaluating, and promoting such resources. Likewise, the technology was designed to be simple and easy to understand by the technical staff, avoiding complex or impractical application processes in the routine practices associated with *ex situ* conservation. Given the need expressed by Latin American technicians during the first workshops on the transfer of the technology developed, the coordination of the program decided to design additional tools to facilitate or make *in situ* conservation processes more efficient.

During the period 2012–2018, 15 workshops were held in 14 countries, eight of them in Latin American countries, three in Europe, one in Africa, and one in Asia. Around 320 technicians from four continents were trained through four regional workshops and 11 national workshops. This way, the Program to Strengthen Capacities in National Plant Genetic Resources Programs in Latin America became a global initiative. Over time, the program became a

technology transfer model for the community of technicians and scientists of plant genetic resources for food and agriculture (PGRFA). This model includes a tool design strategy called CAPFITOGEN (from CAPacidad en recursos FITOGENéticos in Spanish) that helps users to carry out data analysis and processes based on the real and current needs of their routine work. In this model, beneficiaries are guaranteed direct access to scientists and technology developers so that they serve as companions and help in solving their doubts or application difficulties. At the same time, scientists get feedback from the experiences and problems of the technicians who manage agrobiodiversity directly. This informs future technological developments that are even closer to the real needs in national programs. This way, two versions of CAPFITOGEN tools were deployed between 2012 and 2018.

### **1.2.1 Evolution of tools: CAPFITOGEN3 local mode and on server mode**

The process for upgrading the tools began at the end of 2019 with the decision to include a new mode of use in addition to the local installation version and offline use. This new mode presents users with a new version of the tools already installed on a server and website. The process comprised a thorough review of all procedures and the update of functions and R scripts, which resulted in the third version of CAPFITOGEN tools. This was possible thanks to the strong support of the project 'Networking, Partnerships and Tools to Enhance in situ Conservation of European Plant Genetic Resources' (Farmer's Pride), funded by the Horizon 2020 Framework Programme of the European Union, and specifically, the collaboration between the University of Birmingham (UK) and the Universidad Nacional de Colombia.

### **1.2.2 Basic strategy for the design of CAPFITOGEN tools**

The methods to collect, conserve and characterize PGRFA with scientific standards have usually come from regions and centres where economic resources, infrastructure, or staff training do not represent major limitations. As a result, these methodologies cannot be applied in developing countries, or if they are applied, they cannot be extended to all conserved germplasm. This situation contrasts with the fact that the greatest agricultural plant genetic richness is concentrated in developing countries.

This scenario has led some research groups around the world to explore more economic and simpler methodologies for the conservation and use of agrobiodiversity that are much better adapted to the conditions of national programs. The methodological alternatives include the use of environmental information from collection sites (ecogeographic data) to estimate the genetic variability of germplasm or to determine the probability of finding genes of interest in a more successful way. Likewise, the use of geographic information systems (GIS) is proposed to obtain and use the said ecogeographic data. Since most of the ecogeographic information and the software to carry out the analyses are of free access, the investment is reduced to a computer with a commercial set up and training of the staff. Therefore, this technology is compatible with multiple scenarios, including those with limited resources, a condition that often occurs in national programs in developing countries.

Based on the premise that simple but efficient technology has a greater impact and is more adopted by technicians who work daily on the conservation and use of PGRFA, the coordination of the Program to Strengthen Capabilities

in National Plant Genetic Resources realized that this type of technology had to be delivered in the form of simple tools. This simplicity would have to reach the point that future users should be able to be trained in the use of the tools during a short workshop lasting between two and four days.

Since 2012, the tools through which the appropriate technology was introduced to the PGRFA community have been officially disseminated as CAPFITOGEN tools. Their design considered the following premises:

- The tools must be based on scientific advances and contributions previously published in international journals. The tool can fully or partially cover the methodology of that study..
- The selected methodologies must be implemented in the R software (R Core Team, 2020). Some methodologies (as they have been published) have not yet been developed as R scripts. In those cases, if they can be converted into R scripts, they could become CAPFITOGEN tools.
- The methodology is selected for its applicability in processes that are currently carried out in national programs, genebanks, and in situ conservation projects. The methods included in the tools should be simple enough that most PGRFA technicians can understand them and apply them to routine tasks.
- CAPFITOGEN tools, already fully developed as R scripts (programming lines), must be presented to the user in a simple way so that they do not have to learn to use R or its programming language to apply them.
- Finally, CAPFITOGEN tools must be designed in such a way that their delivery and operation do not imply any financial charge to potential users.

CAPFITOGEN tools versions 1 and 2 were introduced within a friendly interface built in HTML and Java script, with the use of a virtual server. The user had to download an installer to install R and Java. Then, the installer opened the internet browser to access the forms to adjust the parameters, configuring the execution of the process (Parra Quijano *et al.*, 2014).

### 1.3. Contents of CAPFITOGEN3 toolbox

For the current version of CAPFITOGEN tools, a complete revision of all R scripts was carried out, updating a high number of functions and processes, and adapting them to the two new ways of use:

1. By downloading the set of R scripts of the tools and the list of parameters that are visualized, configured, and executed in the RStudio Desktop software (<https://rstudio.com/products/rstudio/download/>), distributed under license and terms stipulated by version 3 of the GNU Affero General Public License (<http://www.gnu.org/licenses/agpl-3.0.txt>).
2. By accessing CAPFITOGEN3 tools deployed on a server. the user only enters a website and performs the registration process. Then, the user accesses the tool forms, uploads the tables and information necessary for each process, and requests the execution of the R script on the server. Users can have access to CAPFITOGEN on the server through the link <http://onservercapfitogen.net/>.

The tools developed so far are:

**TesTable (Chapter 3):** Checks the format of tables that will be used in other tools.

**GEOQUAL (Chapter 4):** Facilitates the selection of ecogeographic variables appropriate for the study, which is helpful in the use of tools such as ELCmapas or Modela.

**SelecVar (Chapter 5):** Facilitates the selection of ecogeographic variables appropriate for the study, which is helpful in the use of tools such as ELCmapas or Modela.

**ELCmapas (Chapter 6):** Creates ecogeographic land characterization (ELC) maps which reflect adaptive scenarios for a specific species and country or region.

**Representa (Chapter 7):** Analyses the ecogeographic representativeness of a species within a germplasm collection, detecting biases or any geographic and ecogeographic gap in the target collection.

**ECOGEO (Chapter 8):** Enables ecogeographic characterization of germplasm collection sites. The user has access to more than 170 bioclimatic, edaphic, and geophysical variables.

**DIVmapas (Chapter 9):** Creates maps that show areas of high ecogeographic, phenotypic, and/or genotypic diversity, based on the determination of ecogeographic, phenotypic or genotypic distances in neighbourhoods.

**ColNucleo (Chapter 10):** Facilitates the identification of core collections (subsets) based on ELC maps from proportional allocation strategies. These nuclear-core subsets are representative of the original collection from an ecogeographic point of view.

**FIGS\_R (Chapter 11):** Applies filters to select the most suitable germplasm in relation to abiotic stresses of interest for crop breeders and, thus, generate subsets of 'focused identification of germplasm strategy' or FIGS -type germplasm.

**rLayer (Chapter 12):** Generates sets of ecogeographic layers that do not correspond to the boundaries of a country but to the extent of the distribution of the user's collections/occurrences, which are automatically available for the other applications.

**Complementa (Chapter 13):** Performs a complementarity analysis between cells or protected areas and determines the degree of coverage of current protected area networks in terms of *in situ* conservation of PGRFA.

**Bfuture (Chapter 14):** Facilitates the download and adjustment of bioclimatic information in future adaptation scenarios, which is useful to obtain potential distributions of species in future scenarios for analysis using the Modela tool.

**Modela (Chapter 15):** Creates species distribution models with only the introduction of presence data by the user.

**Mcompare (Chapter 16):** Complementary tool for comparing projections of species distribution in current and future scenarios, determining four possible situations related to their conservation.

**Tzones (Chapter 17):** Facilitates the identification of spatio-temporal seed transfer zones to support tasks to restore vulnerable plant populations in the present and future under climate change scenarios.

## 1.4. Ecogeographic variables

A high number of CAPFITOGEN tools require two basic inputs for their operation: the information provided by the user and a set of environmental variables in the form of GIS layers, which are necessary to carry out ecogeographic analyses. In previous versions, a set of 103 variables was consolidated, including 67 bioclimatic variables, all of them from the WorldClim v 1.4 database (<https://www.worldclim.org/data/v1.4/worldclim14.html>), 31 edaphic variables from the harmonized world soil database HWSD v 1.21 (<https://iiasa.ac.at/web/home/research/researchPrograms/water/HWSD.html>), and five geophysical variables, all derived from the DEM from SRTM v 3 (<http://srtm.csi.cgiar.org/>). For CAPFITOGEN3, an update of the ecogeographic layers was also carried out. The bioclimatic layers were updated to WorldClim v 2.1, including a new monthly variable (vapor pressure) and adding 35 new edaphic variables from original SoilGrids data (<https://soilgrids.org>) assembled to be comparable with the HWSD variables (top and subsoil). Finally, 26 new geophysical variables (solar radiation and wind speed, monthly and annual average) were added from WorldClim v2.1 (<https://www.worldclim.org/data/worldclim21.html#>).







**National Workshop, Sancti Spiritus (Cuba), September 2013.**



## **2 | CAPFITOGEN tools v3: features and functioning**

## 2.1. Local mode

For CAPFITOGEN3, the local (on a desktop PC or laptop) and offline use introduced in the first and second versions of the tools was maintained. However, the user-friendly interface format created in HTML and Java for versions 1 and 2, and that was designed to be viewed in an internet browser connected to a virtual server caused problems to some users. These difficulties required the creation of some patches and temporary solutions.

In order to avoid these type of problems, for CAPFITOGEN3 *local mode*, the user must download the R scripts and provide the list of parameters required by each tool, through an additional R script. The user does not need to know the R software or its programming to run the scripts. To facilitate this contact with the R scripts of each tool and, particularly, the parameter scripts, CAPFITOGEN3 *local mode* relies on software associated with R called RStudio (<http://rstudio.com/>). Therefore, to run CAPFITOGEN3, the user must first download the R scripts of the tools and of the tool parameters, and the required set of folders from the CAPFITOGEN website (<http://capfitogen.net>). Then, the user must download and install R 3.6.3 (<https://cran.r-project.org/bin/windows/base/old/3.6.3/>) and RStudio desktop (from <https://rstudio.com/products/rstudio/download/#download> or directly from <https://download1.rstudio.org/desktop/windows/RStudio-1.3.1073.exe>).

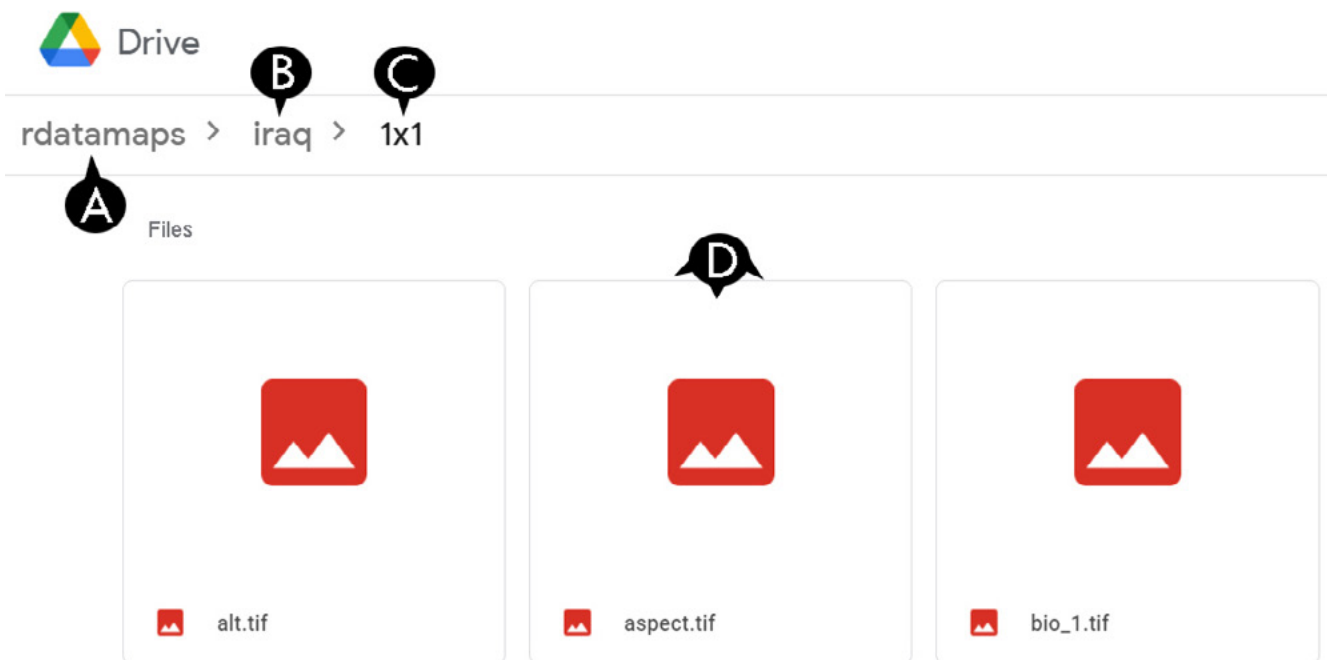
### 2.1.1 How to install CAPFITOGEN3 local mode

The following steps are necessary to install CAPFITOGEN3 *local mode*:

- a) Download the R software for Windows. If you want to perform an analysis with Modela (distribution models based on the biomod2 package), you must download R version 3.1.2 from <https://cran.r-project.org/bin/windows/base/old/3.1.2/>. If you do not intend to use this tool, you can download a newer version of R such as 3.6.3 from <https://cran.r-project.org/bin/windows/base/old/3.6.3/> for greater efficiency. Although CAPFITOGEN3 *local mode* tools have not been tested on macOS, there is a version of R for this operating system. Therefore, in principle, all or almost all CAPFITOGEN3 scripts would be functional in this version of R. The same could be applied to the Linux operating system.
- b) During the R installation process, do not forget to write down the installation path, since you will need it in the next step to access the folder where the software was installed.
- c) Download the R libraries that are necessary for the tools to work. If R version 3.1.2 was downloaded, you must download the libraries from <http://t.ly/GOOF>. If you downloaded R version 3.6.3, the libraries should be downloaded from <http://t.ly/tqu1>. You will obtain a compressed file (.zip) which must be unzipped. As a result, you will get a folder called 'library'.
- d) Replace the original 'library' folder from the R installation (which will be found in the path and folder where R was installed on your computer) with the 'library' folder obtained from the decompression of the .zip file downloaded in the previous step. It is recommended to first delete the 'library' folder from the original installation, and then copy or move the 'library' folder resulting from the decompression process.
- e) Download and install RStudio Desktop software (free version) from <https://rstudio.com/products/rstudio/download/#download>. You can download version 1.3.1073 or a previous version (<http://t.ly/IKTV>).
- f) Download the main set of folders and files necessary for the correct operation of CAPFITOGEN3 *local mode*

from <http://t.ly/F9eq>. You will obtain a file called 'CAPFITOGEN3.zip' of about 1.9 GB. When you unzip this file, a folder called 'CAPFITOGEN3' will appear containing all the files and auxiliary folders necessary to use the tools. You can save this folder anywhere on your hard drive. However, it is recommended that the path that leads to it does not contain any spaces (for example, avoid paths such as 'C:/Mis documentos/CAPFITOGEN3').

g) Then, you need to download the layers of ecogeographic information necessary for the analysis or study which are required by most of the tool scripts. Enter the following link <http://t.ly/89eu> which leads to a shared folder on Google Drive. Fig. 1 shows the set of folders and .tif files of Geographic Information Systems (GIS) layers of ecogeographic information.



**Figure 1.** Levels of GIS layers. A) Upper level, B) country level, C) resolution level, and D) 177 layers. Google Drive allows you to download the information at any of these levels.

In the set of folders and files of CAPFITOGEN3 (downloaded in section f), the path that leads to the downloaded ecogeographic variables is similar to that of the shared folder in Google Drive. The downloaded layers must be located inside a folder with the name that indicates the cell resolution (options: 1x1, 5x5, 10x10, or 20x20). This folder must be inside the country folder (with the name of the country as it appears in Google Drive), which must be inside the rdatamaps folder (located inside the CAPFITOGEN3 folder). For example, if the main folder (CAPFITOGEN3) is located at the root of disk C, and you want to perform an analysis for Argentina with layers with 5x5 km cells using only the altitude layer, then the path must be the following:

**C:/CAPFITOGEN3/rdatamaps/argentina/5x5/alt.tif**

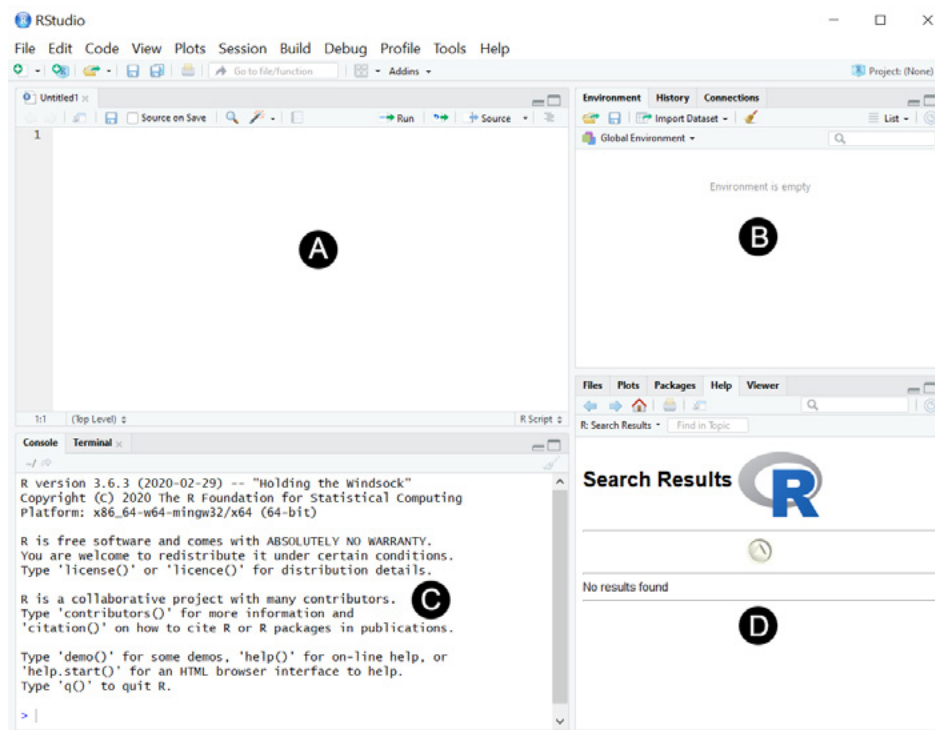
h) Download the scripts for CAPFITOGEN3 tools from <http://t.ly/h7FQ>. You must download the file 'Scripts.zip'; when unzipping this file, the folders 'Parameters' and 'Tools' are generated. In the 'Parameters' folder, you can

find the scripts that allow you to enter the parameters for each tool in RStudio, whereas in the 'Tools' folder you can find the scripts of the tools that allow you to obtain the results. It is not necessary to save these folders and scripts at a specific path.

### **2.1.2 How to use CAPFITOGEN3 local mode**

Once you are clear about the functionality of each tool (described in detail in the following chapters) and have already decided what tool to use, you must look for the parameter script of the tool and double-click on the file 'Parameter\_ToolName\_2020.R'. This script should be opened in RStudio. If the script is not opened, you must set up the system so that RStudio is associated with the .R extension to open these files. If despite configuring the system RStudio does not open when double-clicking on .R files, you must first open RStudio and click on the 'File' tab. Then, select 'Open File' and look for the .R files containing the scripts and parameter list.

Once RStudio has been opened, the following window will appear:



**Figure 2.** Panes of the RStudio desktop v 1.3.1073 software. A) Writing or editing pane. B) Produced objects pane. C) Process progress pane. D) Help pane for R packages, their functions, and graphics visualization.

When the parameter script of a tool such as Complementa is opened, the following writing/editing pane can be observed:

```

1 #####
2 #Parámetros manuales ColNucleo 2020
3 ##### para cada parámetro muestra en color negro; por favor asignar un valor o seleccione una opción
4
5 ruta<-"C:/CAPFITOGEN2"
6 #Parámetro de tipo texto (un texto que va entre comillas "")
7 ##### Nota: Es la ruta donde se encuentra la estructura de carpetas y archivos que son necesarios para ejecutar las herr
8
9 #####
10 #Pasaporte con GEOQUAL
11 #####
12
13 pasaporte<-"Pasaporte/PasaporteOriginalEvaluadoGEOQUAL_ColNucleo.txt"
14 #Parámetro de tipo texto (un texto que va entre comillas "")
15 # Nota: este archivo de texto debe estar en la carpeta Pasaporte, que a su vez es una carpeta que está dentro de "ruta"
16 # Nota: esta tabla es de igual que estructura a otras tablas de pasaporte. Puede haber pasado ya por la herramienta GEO
17
18
19

```

**Figure 3.** Parameter script in the writing/editing pane. A) Tab with the name of the script. B) Title of the script. C) Name of the first parameter. D) You set up the parameter in the area to the left of symbols <— which mean =) according to your conditions or needs. E) The first script line below the parameter line indicates the type of parameter (text, logical, or numeric). This line, like other lines that start with one or several # signs, means that it is only informative. F) The lines ‘# Note:’ (with one or several # signs) offer clues and help the user to set up the parameter correctly. G) Some parameter scripts have sections that are titled in boxes made of # signs.

The types of parameters (text, logical, or numeric) are displayed in Fig. 4.

```

51 #####
52 #Características de la colección núcleo
53 #####
54 #Porcentaje (%) de la colección total que definirá el tamaño (número de entradas) de la colección núcleo/nuclear
55 porcol<-10
56 #Parámetro tipo número (el cual en Rstudio aparecerá con color azul)
57 # Nota: valor que debe estar entre 0 y 100
58
59
60
61
62
63
64
65
66
67
68
69
70

```

**Figure 4.** Types of parameters. A) Subtitle. B) ‘porcol’ numeric type parameter in which the number (in this case 10) is shown in blue. C) Information line indicating that this is a numeric parameter. D) ‘estratcol’ text type parameter, in which the text (in this case ‘P’) is shown in green. E) The options ‘C’ or ‘L’ are also allowed for parameter ‘estratcol’. If you want to use one of these options, you must cancel the first one (‘P’) by adding a # symbol at the beginning of the line (#estratcol<—”P”, which turns the entire line green) and activate the desired line/option by removing the # symbol from the beginning. The name of the parameter turns black when activating the line. Be careful not to leave more than one activated line for a parameter, since only the last activated line will be considered. F) Information line indicating that it is a text parameter. G) Logical type parameter (has only two valid answers, TRUE or FALSE) in which the option is shown in blue (as in the case of the numeric parameter). H) Information line indicating that this is a logical parameter.

In some cases, information may appear in front of a parameter indicating if it must be filled in or adjusted by the user, depending on the configuration of a previous parameter. Fig. 5 shows a fragment of the rLayer tool script in which several parameters must be set up according to what has been specified in parameter 'cropway'. In other cases, a parameter must be set up depending on what has been indicated in a previous parameter or several previous parameters. If you do not have to set up a parameter because a condition has not been met in a previous parameter, you should neither delete the parameter nor modify it.

You must also take some precautions when indicating the path that leads to certain elements within the structure of the hard disk. Each part of a path must be separated with the slash (/) symbol instead of backslash (\), which is a symbol that usually appears in paths in Windows. Also, if you indicate a path, make sure that the element or folder actually exists in R. For example, if you want the results to be saved in the path C:/CAPFITOGEN3/Resultados, you must first make sure that the 'Resultados' folder is already created within the CAPFITOGEN3 folder in the C drive.

```

14 cropway<-"polygon"
15 #cropway<-"square"
16 #cropway<-"buffer"
17 #Parámetro de tipo texto (un texto que va entre comillas "")
18 ##### Nota1: Este parámetro le indica a rlayer como se va a proceder a cortar las capas mundiales (world) para ade
19 ##### Nota2: Si cropway="polygon", se utilizará un archivo vectorial tipo shapefile que proporcionará el usuario q
20 ##### Nota3: Si cropway="square", se utilizarán los sitios de recolección (coordenadas) incluidos en la tabla de p
21 ##### Nota4: Si cropway="buffer", se utilizarán los sitios de recolección (coordenadas) incluidos en la tabla de p
22
23 buffer<-30 #Sólo aplica si cropway="buffer"
24 #Parámetro tipo número (el cual en Rstudio aparecerá con color azul)
25 ##### Nota: Parámetro numérico que expresa kilómetros (km) y con el que se indica el radio que se usará para gener
26
27 shapefile<-"albania" #Sólo aplica si cropway="polygon"
28 #Parámetro de tipo texto (un texto que va entre comillas "")
29 ##### Nota1: En este parámetro se debe indicar el nombre del shapefile, que debe estar en un sistema de coord lat-lon
30 ##### Nota2: El shapefile (los 4-7 archivos que lo componen, entre los que debe estar los imprescindibles .shp, .dl
31
32 pasaporte<-"Pasaporte/PasaporteOriginalEvaluadoGEOQUAL.txt" #Sólo aplica si cropway="square" o cropway="buffer"
33 #Parámetro de tipo texto (un texto que va entre comillas "")
34 # Nota1: este archivo de texto debe estar en la carpeta Pasaporte, que a su vez es una carpeta que está dentro de
35 # Nota2: esta tabla es de igual que estructura a otras tablas de pasaporte. Puede haber pasado ya por la herramie

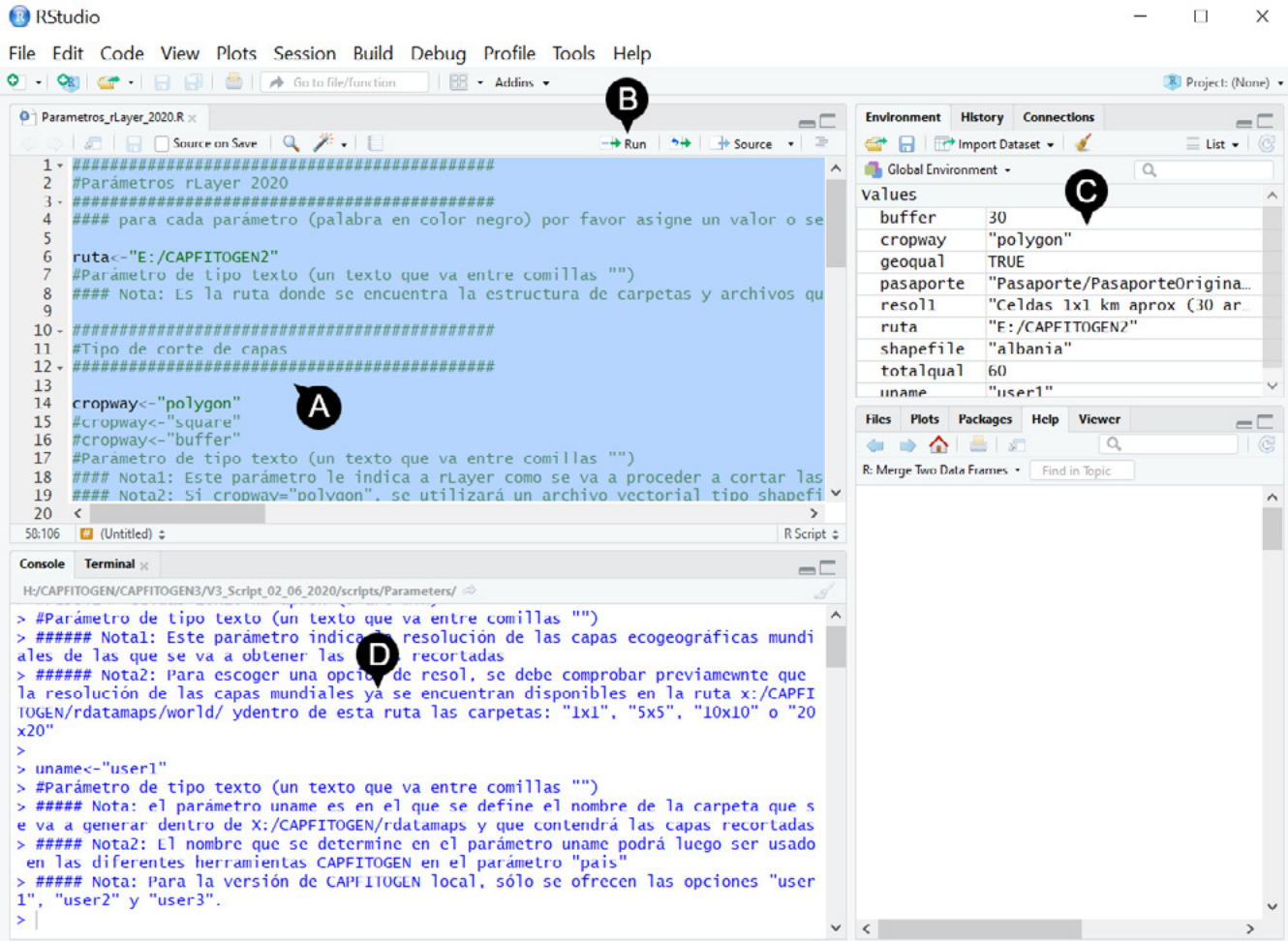
```

**Figure 5.** Parameters that depend on other parameters (using the 'rLayer' script as an example). A) Parameter 'cropway' (reference for other parameters). B) Parameter 'buffer' which must be adjusted if 'cropway' has been previously set up as 'buffer'. C) Parameter 'shapefile' in which you must provide the name of a shapefile-type file (with extensions .shp, .shx, .dbf, etc.); this parameter is applicable only if 'cropway' is set up as 'polygon'. D) Parameter 'pasaporte' which must be adjusted according to the user's conditions only if 'cropway' is equal to 'square' or 'buffer'.

Adjust only the parameters that apply to the case or to the use that you want to make of the tool; the rest can be left as they appear in the script. Once you have finished adjusting the parameter script according to your conditions and available data, you must select all the lines of the script and execute it, as shown in Fig. 6.

Before running the script of the corresponding tool, you should check in the upper-right pane if the elements you created are correct. If you find that you have made a mistake, you can easily fix the problem by correcting the parameter script and repeating the steps to run it again.





**Figure 6.** Execution of the parameter script of a tool, in this case, rLayer. A) After setting up the script, select it completely, so that the upper-left pane is highlighted in blue. B) Run the script by clicking on the little button that says 'Run'. C) After running the script, a list of objects that have just been created and that will be required by the tool's script will appear in the upper-right pane. D) Blue text indicating the actions carried out will also appear in the process progress pane (lower-left window). In the event of an error, a red text message will appear in this pane.

If you want to save the parameter script settings for future procedures, simply go to the 'File' tab and select the option 'Save as...'. This way you can save the script with the desired configuration and then run it whenever you want.

After opening the tool script, it will appear in the upper-left pane and it will be executed in the same way as the parameter script (see Fig. 6). The tool script is usually quite long, so you must be careful when selecting it since it must be completely highlighted in blue (background colour) from the first to the last line. Then, you must click on the 'Run' button, and every process that is being run will be gradually shown in blue text in the lower-left pane. When execution is finished, the generated elements will appear in the top-right pane and you will be able to search for the results in the designated folder.

You must take certain precautions when executing the tool script that corresponds to the parameter script that has been previously set up and executed. Multiple errors will occur if there is no match between these two scripts. On the other hand, the tool script should NOT be modified unless you are an expert in the R language and want to introduce some variation on the original script. Even with some expertise, a voluntary or unintended modification on the original script can generate unexpected or unwanted products, block the execution of functions, and prevent the generation of any product.

### 2.1.3 Errors in CAPFITOGEN3 local mode

As indicated above, errors that may occur in the execution of both the parameter script and the tool script will appear in red text in the process progress pane (lower-left window).

```

Source
Console Terminal x
> library(dismo)
> library(rgdal)
rgdal: version: 1.4-8, (SVN revision 845)
Geospatial Data Abstraction Library extensions to R successfully loaded
Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
Path to GDAL shared files: C:/Users/hmparraq/Documents/R/win-library/3.6/rgdal/gdal
GDAL binary built with GEOS: TRUE
Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
Path to PROJ.4 shared files: C:/Users/hmparraq/Documents/R/win-library/3.6/rgdal/proj
Linking to sp version: 1.4-1
>
> #introducción tabla de lista de países y resoluciones de extracción a elegir y traducción
> #Rversion
> vvv<-R.Version()
> vvv<-as.numeric(vvv$year)
>
> if(vvv<=2019){
+   resol<-read.delim("resol.txt")
+ }
> if(vvv>2019){
+   load("resol.Rdata")
+ }
Error in readChar(con, 5L, useBytes = TRUE) : cannot open the connection
In addition: Warning message:
In readChar(con, 5L, useBytes = TRUE) :
cannot open compressed file 'resol.Rdata', probable reason 'No such file or directory'
> resol<-subset(resol,resolucion==paste(resol1))
Error in subset(resol, resolucion == paste(resol1)) :
object 'resol' not found
> resol<-as.character(resol[1,2])
> setwd(paste(ruta,"rdatamaps",sep=""))
> dir.create(as.vector(paste(uname)))
Warning message:
In dir.create(as.vector(paste(uname))) : 'user2' already exists
> setwd(paste(ruta,"rdatamaps/",uname,sep=""))
> dir.create(as.vector(paste(resol)))
Warning message:
In dir.create(as.vector(paste(resol))) : '1x1' already exists
> setwd(paste(ruta))

```

**Figure 7.** Text in red that may appear in the process progress pane. A) Indicates that R was requested to load some packages (dismo and rgdal). B) R shows (in red) that the packages have been uploaded for the first time, which is NOT an error. A new upload will not generate any text in red. C) Here, a true error message is displayed, which always starts with the word 'Error'.

The message then reads '...cannot open the connection...' and later says "cannot open compressed file 'resol.Rdata'". This means that R cannot find the file 'resol.Rdata' where it should be within the CAPFITOGEN3 set of folders and files, which was probably caused by a misconfiguration of a parameter. D) This second error message warns the user that object 'resol' was not found. This second error is a consequence of the first error since it was not possible to introduce a table, and this did not allow the creation of an object in R. Therefore, the first error to correct is that of message C). E) Sometimes, the red text indicates a warning about something that could be irregular or that the person who designed the function in R considers that the user should know. This does not correspond to an error.

Messages in red can be due to various causes, not necessarily errors; however, all errors will appear in this colour. As shown in Fig. 7, loading R packages for the first time produces a series of text in red that in most cases does not offer information of greater importance to the user. Additionally, text in red can appear because R wants to warn the user of something, but this cannot be considered an error. These messages are preceded by the phrase ‘Warning message:’ and then indicate the subject to be reported. Sometimes, when warning messages are numerous (as they come from repetitive processes), it is necessary to write `warnings()` in the last line at the bottom of the text in the process progress pane. Then, click on ‘enter’ and the warning messages will appear afterward. Finally, some error messages can be found, such as the examples shown in Fig. 7.

Text in black can also appear indicating an action that has been carried out, but that would not correspond to an error. Several of the most frequent errors are listed in the final part of this user manual, along with an explanation of why they can occur and the alternative solutions.

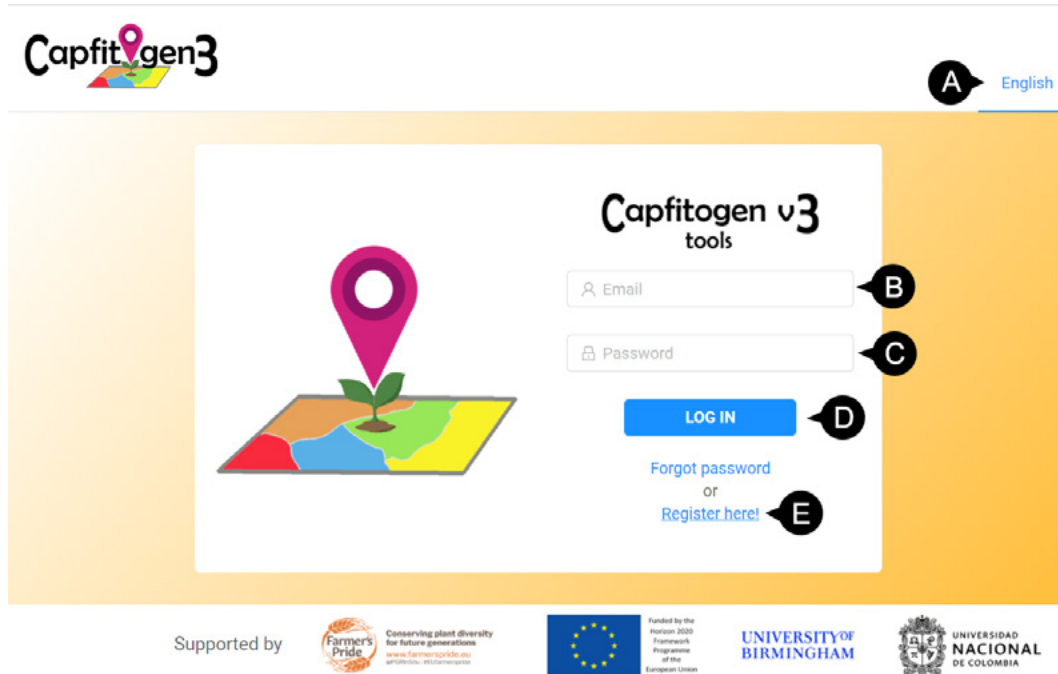
## 2.2. On server (online) mode

The third version of CAPFITOGEN tools introduces on server mode, an important novelty to users. In this new mode, no installation process or specific software is required to use the tools. Additionally, you do not need to have a computer with special requirements to obtain the desired products. Simply go to CAPFITOGEN3 URL using your browser of choice and register as a user to have access to the tools and produce the results you desire. The R software and scripts are already installed on the server and are displayed and configured through user-friendly forms. After setting up the operation of the tool, you can start the analysis and wait for the results to be saved in your user area. This is a private space that stores both the user’s particular tables necessary for the analyses and the results of the application of the different tools.

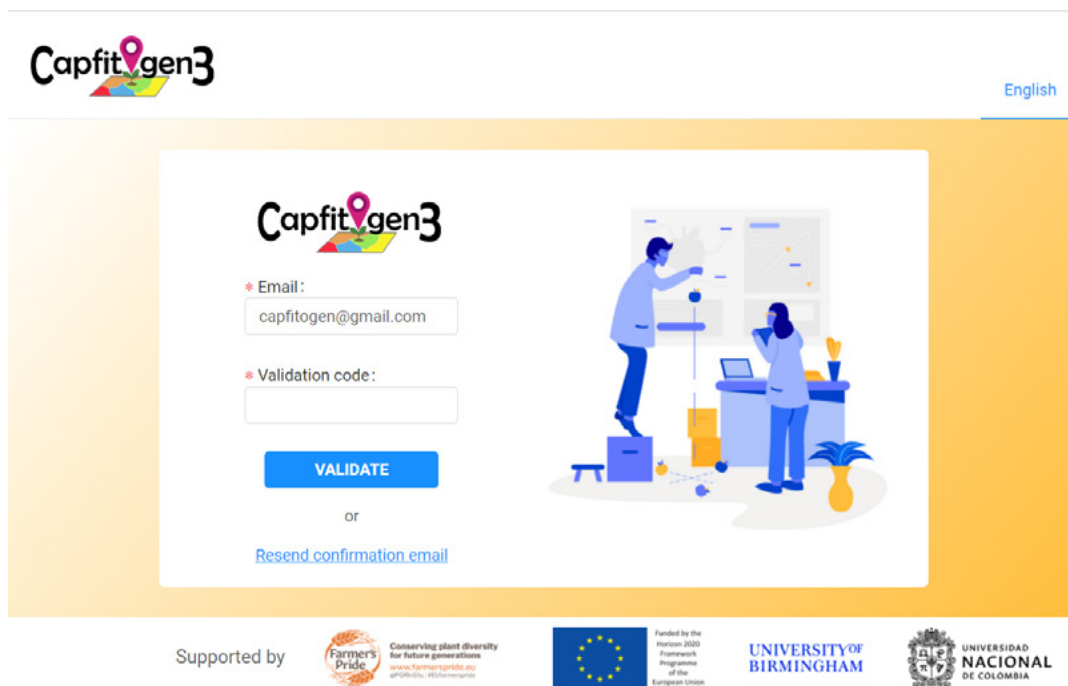
### 2.2.1 How to register in CAPFITOGEN3 on server mode

The procedure for registering as a CAPFITOGEN3 user is very simple and similar to the registration process on almost any platform, using validation via email. The steps are:

1. Go to <http://onserver.capfitogen.net/>. An access page such as that shown in Fig. 8 will be displayed. Start your registration process by clicking on ‘Register here’ (Fig. 8E). A form will appear which must be filled out with the following information: name, last name, email, and password (which should be confirmed in an additional field in the form). Finally, tick the box that appears to accept the terms and conditions and click on the ‘Register’ button to finish the registration.
2. You will get an email from [no-reply@verificationemail.com](mailto:no-reply@verificationemail.com) with a code that needs to be memorized. You do not need to reply to this message.
3. The registration page leads to another page to validate the code you have just received by email (Fig. 9). There, you must indicate the email you registered and the code received. Then, click on the ‘Validate’ button.



**Figure 8.** Entry portal to CAPFITOGEN3 on server mode. A) Language settings. In this part, you can select the language between Spanish and English. B) Space to enter the email the user registered with. C) Space to enter the password of the registered user. D) Button to log in as a registered user. E) Link to start the registration process.

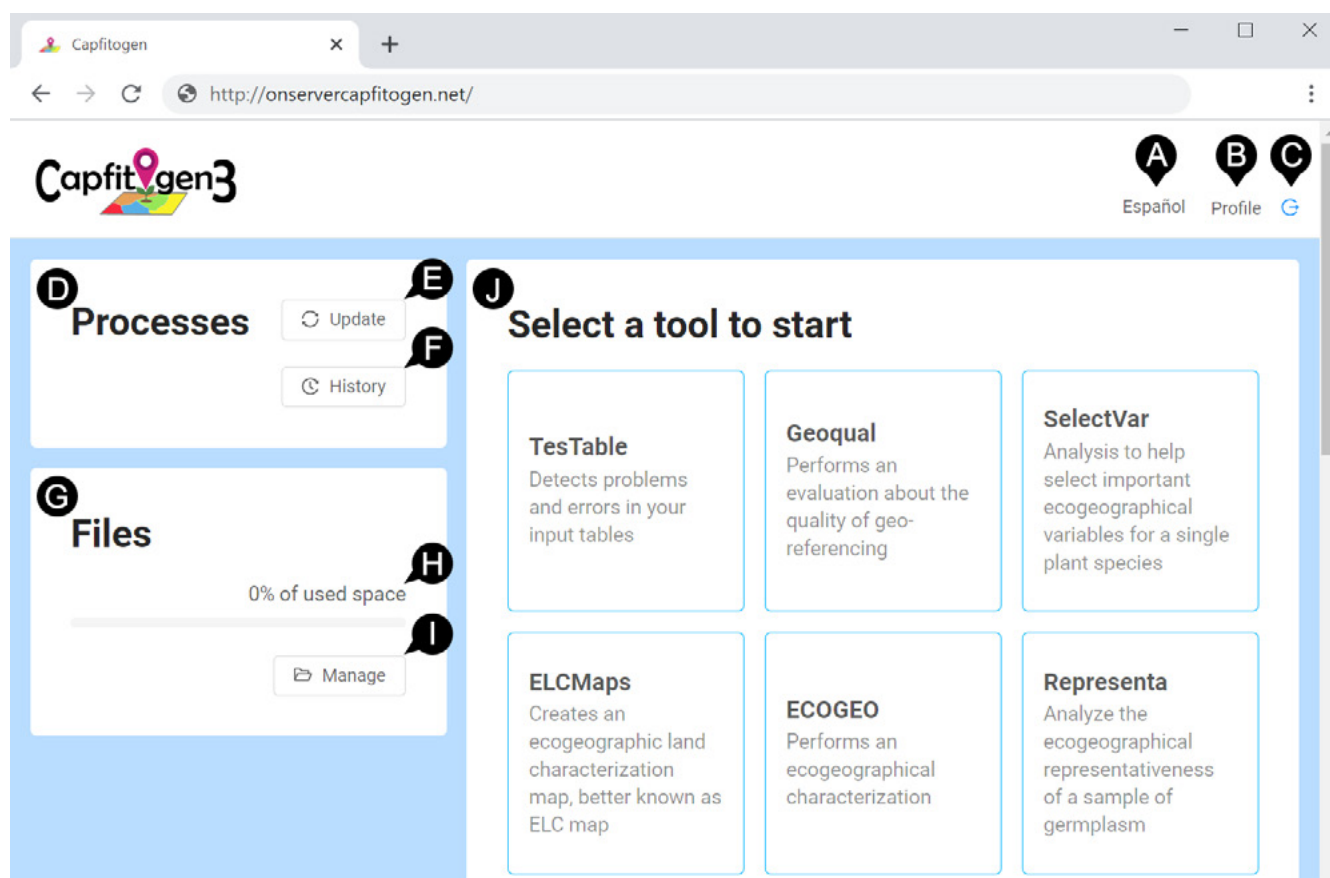


**Figure 9.** Registration validation page. On this page, the user enters the code received via email.

4. Once the code has been successfully validated, you can now log in using the email address and password defined in the registration process. To do this, the access page to CAPFITOGEN3 on server is automatically displayed. If this does not happen, you can simply go to <http://onserver.capfitogen.net/> using your browser of choice and the access page will be available.

### 2.2.2 How to use CAPFITOGEN3 on server mode

After logging in, the main view of the tools will open. Different panels will be displayed, as shown in Fig. 10.

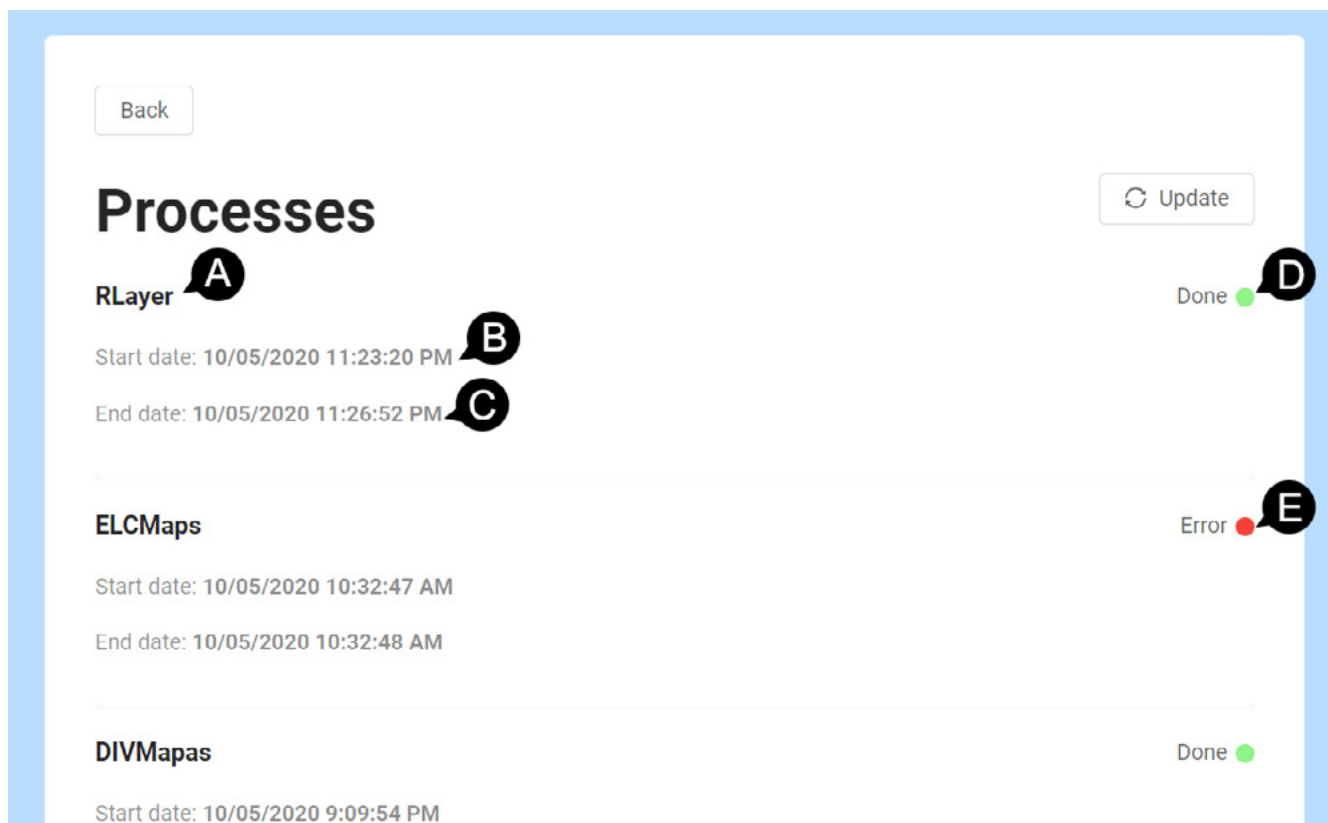


**Figure 10.** Main view of CAPFITOGEN3 on server mode tools. A) Language selection. B) Access to user profile information. C) Logout button. D) Process area. E) Button to update the status of processes. F) Button to access the history of the processes executed. G) User's Files and Results area. H) Indicator of the percentage of space assigned to a user that is already in use. I) Button to access the files stored in the space assigned to the user. J) Tool selection panel.

You have four sections to perform actions:

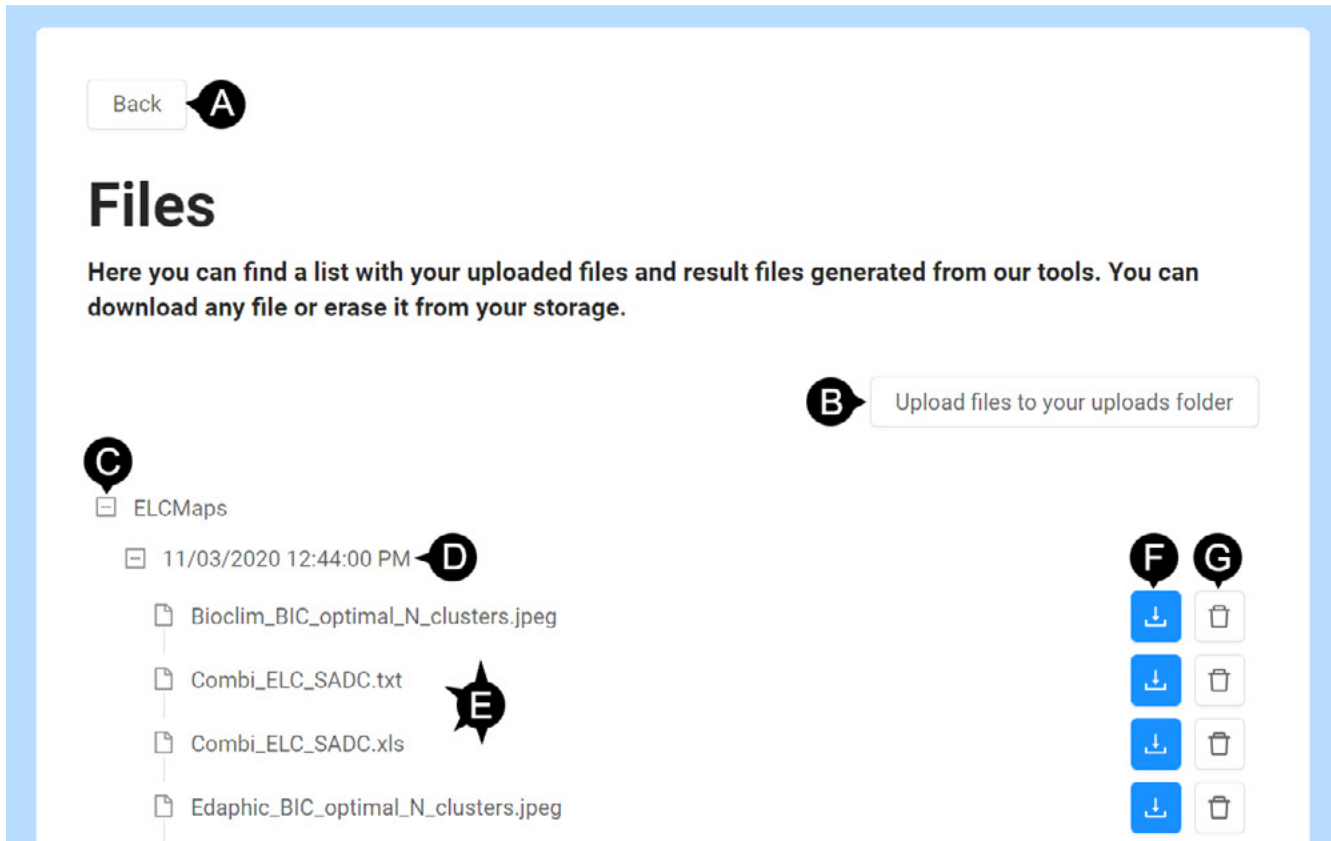
1. Tool selection panel: Here you can search for the tool you want to work with. When clicking on the name of the tool of interest, a blank form of the tool is displayed (Fig. 10-J).

2. Process area: This area is made up of the panel that shows the progress of a tool's execution (Fig. 10-D), and the history which is accessed through the button indicated in Fig. 10-F. The sections of the process area are detailed in Fig. 11. Traffic light colours (green, red, or yellow) indicate if the process has finished or is still running.



**Figure 11.** Process area. A) Name of the tool that was run. B) Date and time the tool analysis started (in this case, RLayer). C) Date and time the analysis was completed. D) Status of the process. In this case, a green circle reports that the analysis has been executed successfully and the word 'done' confirms that the process has already been carried out. E) In this case, an error is reported (with a red circle), and therefore, the process did not end correctly. A yellow report may also appear, indicating that the process is still pending completion.

3. User's Files area: This is the space assigned to each user for the storage of your work files (passport tables, external source tables, characterization tables, etc.) or the results of the execution of tools. Fig. 12 shows the interior of the User's Files area which is accessed through the button indicated in part I of Fig. 10. In this area, the results of the analyses carried out by the tools are saved in folders with the name of the tool they come from. The results for the same tool are separated into subfolders by the day and time they were obtained. The result files are within these subfolders and can be downloaded and/or deleted to free up space. In this area, you can also find a folder with the name 'uploads' which contains all the files that the user has uploaded to carry out the processes of the tools. These files can be uploaded by clicking on the button 'Upload files to your uploads folder' (Fig. 12-B) or the buttons in each tool enabled for the you to upload files.



**Figure 12.** User's Files and Results area. A) Button to go back and return to the main view. B) Button to upload files necessary for the execution of tools. If you upload the files in advance, it is not necessary to upload them again when accessing the tool form; they can be selected from a drop-down list. C) Folder with the name of the tool that produced the results. D) Subfolder with the information on the day and time of execution of the tool that produced the results. E) Files resulting from the execution of the tool. F) Button to download the corresponding result file. G) Button to delete the corresponding file from the User's Files and Results area.

4. User's Profile Area: Accessible through the link shown in Fig. 10-B. In this area, you can modify the information requested in the registration process (Fig. 13).


The forms to configure the parameters of each tool, which are accessible through the tool selection panel, are used similarly to the CAPFITOGEN version 2.0 forms. However, some resources appear in on server mode. It is important to note that in CAPFITOGEN3 *on server mode*, the parameters 'ruta' and 'resultados' disappear since they are no longer necessary. Also, the parameters to enter the names of table or map files are now buttons to select such files and upload them to the server. However, you can upload those files and then, when accessing the form, simply select the appropriate file from a list of uploaded and stored files in the User's Files and Results area. These files are located inside the 'uploads' folder.

[Back](#)

## Mauricio Parra Quijano

hmparraq@unal.edu.co

[change password](#)



To change your personal information please use the form below

\* Name:

\* Last name:

Title:

Institute or company:

What you will use the tools for:

Secondary email:

Receive notifications from our newsletter

[Update](#)

Figure 13. User's profile and password change area.



Fig. 14 shows the tool form layout and the different types of parameters that can be found in on server mode:

The screenshot shows the 'TesTable' tool form. At the top left is a 'Back' button. At the top right is a yellow 'Download manual' button. The title 'TesTable' is prominently displayed. Below the title is a brief description: 'This tool was create to detect problems and errors in your input tables'. The form contains several input fields: a dropdown menu for 'Table type' (with a description 'tiptable. Type of table to be analyzed'), a file selection dropdown (with a description 'pasaporte. Upload or select file with table to be analyzed and / or corrected'), an 'Upload table' button, a numeric input field for 'access' (with a description 'access. Number of accessions' and a value of 1), and a checkbox for 'fixthem' (with a description 'Select this option to create a corrected table with the detected errors automatically fixed (as much as possible), if this option is false you'll get a report with all the errors'). At the bottom is a blue 'START' button.

**Figure 14.** Example of a tool form (TesTable in this case). A) Name of the tool. B) Button to download the user manual (particularly the chapter for the tool). C) Brief description of the function of the tool. D) The fields to set up the parameters are on the left. E) Name of the parameter (first element on the right). F) A brief description of the parameter can be found after its name. G) For parameters such as 'pasaporte', which require the user to upload a file, two options appear; a drop-down list (to select a file that was previously uploaded and kept in the 'uploads' folder) or a button that allows searching for the file on the local machine and then uploading it to the server. H) Numeric type parameter, where the number that configures the parameter is increased or decreased. I) Checkbox type parameter, where you can tick a box and a tick symbol (✓) appears. This is equivalent to answering the question of the parameter as 'yes' or TRUE. Leaving the checkbox unticked means saying 'no' or FALSE. J) Button to start the process or execution of the tool, once all the parameters have been set up.

Cropway  ^

Square

Buffer

Polygon

Map name

START

cropway. This parameter indicates how the world layers will be cut to adapt them to certain limits

buffer. Choose the radius of the circular area of influence or neighborhood (in km). This area is created on the basis of each cell centroid on the map showing collection sites and clusters generated from accessions whose collection sites are included. The value of the indexes and average distances of each cluster will be assigned to the cell from whose centroid the area of influence was drawn

uname. Select a name for the resultant merged layers

**Figure 15.** Drop-down list of options and text parameters. A) Drop-down list parameter, where only one option can be selected. B) Text parameter, where the user must enter a text (usually short and without spaces).

SADC  ^

10x10 km cell size ~(5 arc-min)  ^

Annual mean temp × Annual prec ×

Prec seasonality × |

Annual mean temp ✓

Mean temp warmest quarter

Mean temp coldest quarter

Annual prec ✓

Prec wettest month

Prec driest month

Prec seasonality ✓

Prec wettest quarter

pais. Select the country/region where all or most of the data accessions you wish to analyze were collected

resol1. Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution but requires greater computing capacity and takes far longer than 5x5 km, particularly in countries with a large land mass. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents or continents

bioclimv. Select the bioclimatic variables you wish to analyze. Vapour pressure is only available if you have previously downloaded this variable using Wclim2 tool

edaphv. Select the soil variables you wish to analyze

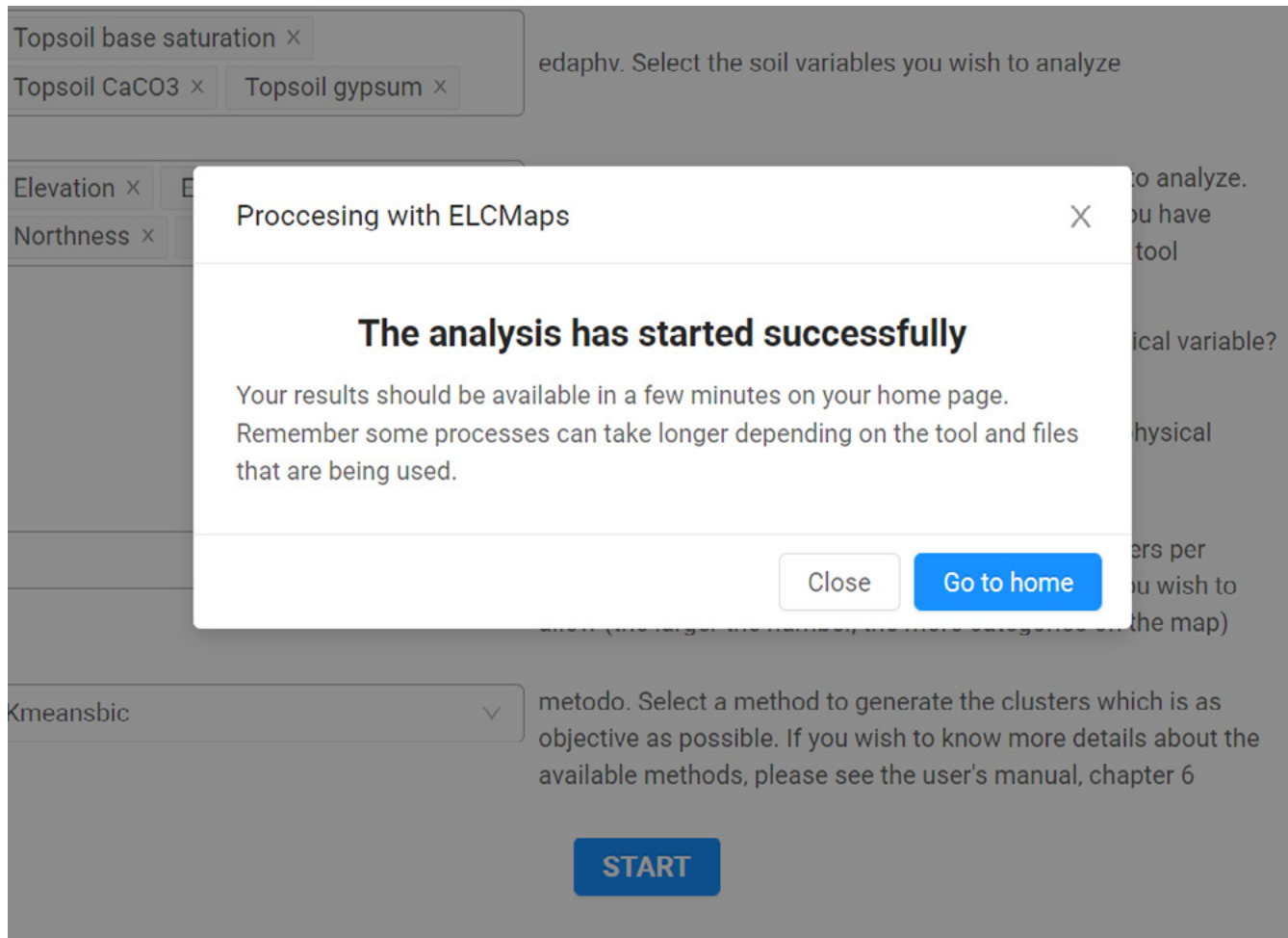
geophysv. Select the geophysical variables you wish to analyze. Wind speed and solar radiation are only available if you have previously downloaded these variables using Wclim2 tool

latitude. Do you wish to include latitude as a geophysical variable?

longitude. Do you wish to include longitude as a geophysical variable?

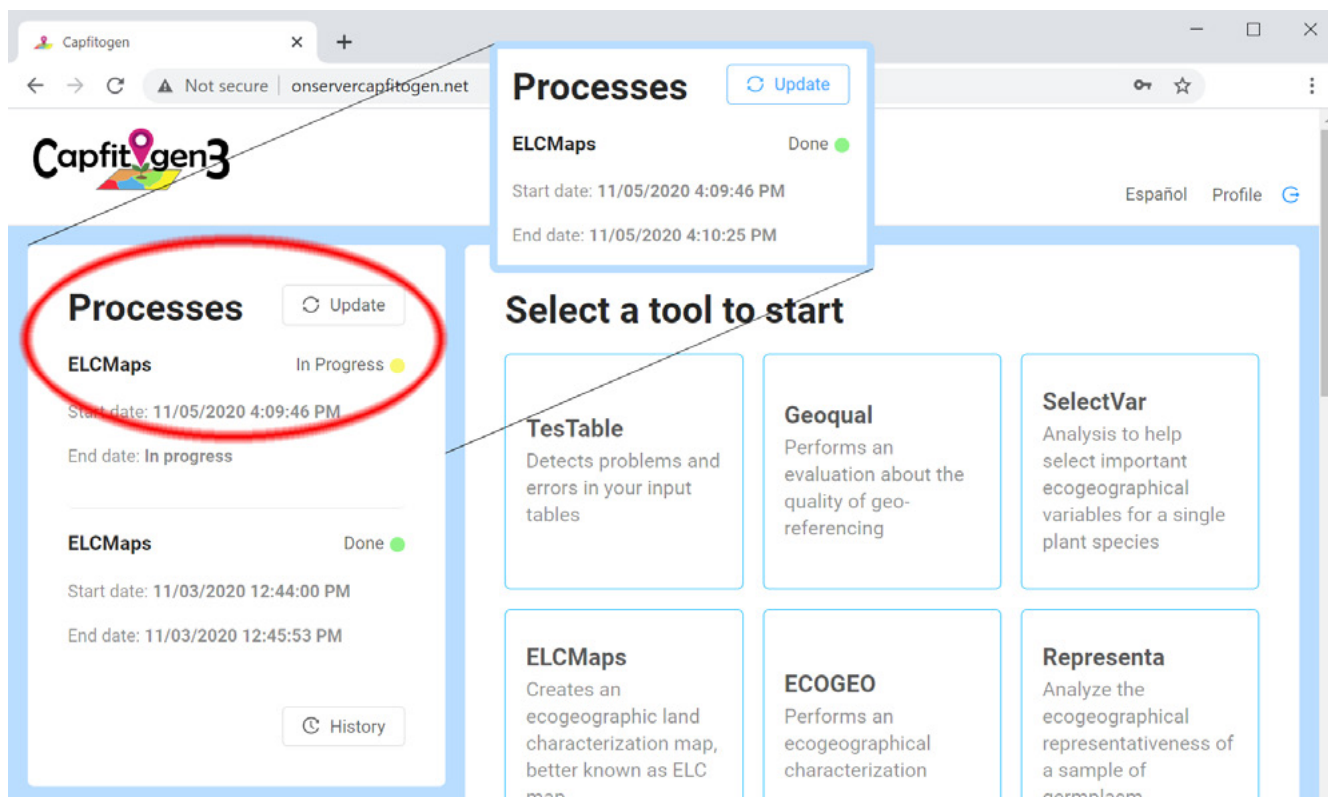
**Figure 16.** Other types of parameters in CAPFITOGEN3 on server mode. A) Drop-down list parameters, in which only one option can be selected. B) Drop-down list parameter with multiple selections. C) The selected options appear inside the box and can be removed from it by clicking on the X. D) On the drop-down list, you can select the variables you wish to include by clicking on them.

After setting up the parameters of the tool and clicking on the button to start the process (Fig. 14-J), a window will appear notifying the user that the analysis has started in R and that the results will be available in a few minutes (Fig. 17).



**Figure 17.** Notification that the process has successfully started. The message also indicates that the results will be available in the main view (particularly in the User's Files and Results area) depending on the complexity of the process carried out by the tool and the size of the files to be processed.

You can go back to the main view where you may find a yellow circle (with the legend 'in progress') in the process area indicating that the results are still to be obtained. This may depend on the time the tool requires to carry out the process. The user can even exit the system and come back later to check if the results are already available. When the results are available, a green circle will appear in front of the process. The user can also wait there and, from time to time, click on the 'Update' button in the process area until the circle turns green and the word 'Done' appears. Then, it is possible to enter the User's Files and Results area to download the results (Fig. 18).



**Figure 18.** Main view after returning from a tool form and starting its process. In the process area (highlighted in a red circle) the last process executed appears at the top, and the previous one(s) in the timeline down. In this case, a yellow circle and the legend 'In Progress' indicate that the results have not yet been produced. After a reasonable period, you can click on the 'Update' button. If the process has finished, the circle will turn green (successful completion), and the legend 'Done' will appear as illustrated in the image.

An 'Error' status with a red circle appears in the process area when the process does not finish successfully and an error occurs (Fig. 19-A). To know the nature of the error, it is necessary to enter the User's Files and Results area. A file named 'ERROR.txt' will appear within the date and time of execution subfolder located inside the tool folder (Fig. 19-B). If you want to find out about the error that has occurred, it is necessary to download the text file and then open it with any text viewing or editing software. The description of the error can be compared with the errors typified in Chapter 18 'Frequent errors' to solve the problem.

The screenshot displays two main panels: 'Processes' and 'Files'.

**Processes Panel:**

- Geoqual:** Status 'Error' (red circle with 'A'). Start date: 11/05/2020 7:10:48 PM, End date: 11/05/2020 7:11:11 PM.
- TestTable:** Status 'Done' (green circle). Start date: 11/05/2020 7:10:01 PM, End date: 11/05/2020 7:10:02 PM.
- ELCMaps:** Status 'Done' (green circle). Start date: 11/05/2020 4:09:46 PM, End date: 11/05/2020 4:10:25 PM.

**Files Panel:**

- Header: 'Here you can find a list with your uploaded files and result files generated from our tools. You can download any file or erase it from your storage.'
- Button: 'Upload files to your uploads folder'
- Folder structure:
  - ELCMaps
  - Geoqual
    - 11/05/2020 7:10:48 PM
      - ERROR.txt** (marked with 'B') - Includes a blue download button (marked with 'C') and a trash icon.
      - Error\_process\_info.txt - Includes a blue download button and a trash icon.
      - Parameters\_Geoqual.txt - Includes a blue download button and a trash icon.

**Figure 19.** Notification of errors in the execution of a CAPFITOGEN3 tool on server mode. A) Error message in the process area, with an 'Error' legend and a red circle. B) In the User's Files and Results area, a file called 'ERROR.txt.' appears inside the execution date and time subfolder located within the tool folder. C) To see the error message produced by R, the file 'ERROR.txt' can be downloaded by clicking on this blue button.



National Workshop, Tirana (Albania), November 2014.



# 3 | TestTable Tool

## 3.1. Data table formats that are accepted by CAPFITOGEN3 tools

Examples and templates of the different types of tables used in CAPFITOGEN3 are available at <https://drive.google.com/drive/folders/1xCnllZgzW0uDeClDvcxbADv9H583xzpn?usp=sharing>. The user will be required to include some of the following data tables according to the tool to be used:

### **3.1.1 Passport data table**

This table contains the fields indicated on FAO-Bioversity's multi-crop descriptors list (MCPD format onwards) version 2 published in 2012. Five fields have been added to this list related to administrative locality description (ADM1, ADM2, ADM3, and ADM4) and country name (not the ISO code) (NAMECTY). Although different formats of passport data tables are used in different countries or research institutes, the MCPD descriptor list (either version 1 or 2) has been the reference for many genebanks and information systems (such as Genesys - <https://www.genesys-pgr.org/>) for storing passport data, usually with some minor modifications and adjustments.

The standard passport data table used in CAPFITOGEN contains 45 fields (40 belonging to the 2012 MCPD descriptor list, plus 5 extra). However, more fields may appear; for example, if the user includes a field related to the availability of the germplasm (field AVAILAB) which is very useful for ColNucleo or FIGS\_R tools. Also, if GEOQUAL tool has been run, which determines the quality of the collecting sites georeferencing, the table should contain the following 5 additional fields: SUITQUAL, LOCALQUAL, COORQUAL, TOTALQUAL & TOTALQUAL100. All the additional fields regarding the inclusion of information about germplasm availability or the quality of georeferencing will be found after the 45 standard fields (on the right side of the table).

### **3.1.2 External source data table**

This table contains 15 fields, some of them matching the name and content of the passport data table. This table has been designed to insert into Representa tool population occurrence data from external sources different from the genebank, or inventories of target plant genetic resources (information that is already included in the passport data table). No additional fields are allowed in this table.

### **3.1.3 Phenotypic data table**

This table contains a column for accession identification (ACCENUMB). The information here should correspond with the identification field of the passport data table. It also includes two or more fields for the phenotypic variables. Both, quantitative and qualitative variables are allowed in this table as long as they show numeric data.



### **3.1.4 Type of phenotypic variables data table**

As the above phenotypic data table allows quantitative and qualitative variables, the type of variable should be specified here. This table only contains three fields:

**ID:** an identification number; it could be consecutive.

**NOMVAR:** the name of each variable as shown in the phenotypic data table (following the same order).

**NATVAR:** here the type of variable is specified. The options are as follows (without stress marks): “Cuantitativo” (quantitative), ‘Nominal’ (nominal), ‘Ordinal’ (ordinal), ‘Binaria simetrico’ (symmetric binary), and ‘Binario asimetrico’ (asymmetric binary).

### **3.1.5 Genotypic data table**

The first column in this table is for accession identification (ACCENUMB). The information here should match the identification field of the passport data table. The table also includes two or more fields for the genotypic variables. Genotypic variables correspond to molecular characterization data using dominant or co-dominant DNA markers, shown as presence/absence of the band. Therefore, this table only allows non-symmetric/asymmetric binary variables, coded 0 (absence) or 1 (presence).

Please note that blank cells are not allowed in any of the tables used by CAPFITOGEN tools. In case there is information missing regarding an entry, variable, or field, fill in the blank cell with the text ‘NA’ (which stands for ‘not applicable’).

CAPFITOGEN requires the tables to be converted into a tab-delimited file (.txt file extension). To facilitate users the task of adjusting data to the required format to be used by CAPFITOGEN tools, a set of Excel files containing those fields needed for each type of table is provided within the ‘Formatos\_Formats’ folder. Some of the Excel tables are given with a few data entry examples (fake accessions) to illustrate how to correctly insert the information. In case some help is needed regarding either the meaning of the fields or how to adequately insert the data, simply position the mouse pointer over the column heading. A window will appear with a brief explanatory comment (in Spanish or English) giving instructions on what the field should contain and how to fill it in (Fig. 20).

	X	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	LATITUDE	DECLONGITU	LONGITU		OREFMETH	ELEVATION	COLLDATE	BREDCODE	BREDNAME	SAMPST
2	4124--N	NA	002094		NA	NA	19780301	NA	NA	
3	452556S	NA	0684645W		NA	120	1999----	NA	NA	NA
4	00----S	NA	079----W		NA	NA	NA	NA	NA	
5	103936N	NA	0671051W		NA	14	NA	NA	NA	
6	252712N	NA	1112853W		NA	0	201002--	NA	NA	NA
7	NA	NA	NA		NA	NA	194812--	NA	NA	
8	412403N	2.1628	0020945E		NA	NA	19780301	NA	NA	
9	NA	-68.7717	NA		NA	120	1999----	NA	NA	NA
10	NA	-79	NA		NA	NA	NA	NA	NA	
11	NA	-67.202	NA		NA	14	NA	NA	NA	
12	NA	-111.485	NA		NA	0	201002--	NA	NA	NA
13	NA	NA	NA		NA	NA	194812--	NA	NA	
14										

**Figure 20.** Image capture of the model passport table showing the field headings. A) Highly important field (green). B) Fields that are not important for analysis with CAPFITOGEN tools, without colour. C) Fields that may be useful for some CAPFITOGEN tools, in yellow. D) By placing the mouse pointer over the column heading in the Excel tables, a comment with directions about the descriptor will appear.

The Excel tables of phenotypic data, type of phenotypic variables data, and genotypic data include an additional spreadsheet that explains in detail how to fill in the table (Fig. 21).

	A	B	C	D	E	F	G
1	ACCENUMB	D1	D2	D3			
2	acc_001	3910	1	0			
3	acc_002	3304	5	0			
4	acc_003	4728	4	0			
5	acc_004	903	5	0			
6	acc_005	3284	5	0			
7	acc_006	914	5	1			
8	acc_007	504	5	0			
9	acc_008	723	2	0			
10	acc_009	4402	3	1			
11	acc_010	3978	2	1			
12	acc_011	267	4	0			
13	acc_012	2942	1	0			
14	acc_013	710	5	1			
15	acc_014	3492	1	1			
16	acc_015	4236	4	0			
17	acc_016	3960	3	1			
18	acc_017	4513	3	1			
19	acc_018	184	3	1			

**Figure 21.** Image capture of the phenotypic data table format where two spreadsheets can be observed. A) In the 'Morphology' or 'Phenotypic' spreadsheet (where this type of information must be included), two types of columns appear; the first is the one reserved for the accession code (ACCENUMB), which is also found in other tables. This column is green and of high importance. B) On the same sheet, other columns without colour can be found. These columns are also important since the characterization information should be included here. C) 'Observations' or 'Directions' spreadsheets, where you can consult the instructions for filling part A.

Once the user has completed the Excel table formats with their data, they should be saved as both, Excel files (.xls or .xlsx) and tab-delimited text file (.txt, Excel gives this option when exporting data). The first Excel extension will allow possible further editing, and the second one (.txt) corresponds to the format accepted by CAPFITOGEN tools.

## 3.2. What are the features of TesTable?

TesTable tool verifies that the tables completed by the user meet the requirements for CAPFITOGEN tools to adequately process them. TesTable can analyse all the tables defined in section 3.1 above.

In TesTable, the user inserts the tables indicating the main characteristics. The tool automatically generates a report with details of all the errors found when meeting the required standard format. At this stage, the user can choose the table to be corrected automatically and a new table without errors ready to be used in CAPFITOGEN tools will be generated. As TesTable cannot assume data without enough information (e.g., coordinates introduced in the wrong format), the text 'NA' will be displayed in those cells where errors were found in the corrected table. It is important to highlight here that according to the type of error found, a manual correction rather than the automatic display of NA values may be a better option.

## 3.3 Using TesTable tool

Once the parameter and TesTable tool scripts have been loaded in RStudio or said tool has been selected in *on server mode*, it will be necessary to define a series of parameters for the R programming to work correctly. The tables to be analysed and optionally corrected by TesTable must be located in the 'Pasaporte' folder.

After defining all the parameters and paths that TesTable needs, click on 'Run' (*local mode*) or 'Start' (*on server mode*) and the analysis process will start.

In the case of TesTable, the results will be produced after a usually short period of time. Such results will be saved in the same 'Pasaporte' folder or in the User's Files and Results area for *on server mode*.

### **3.3.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) to indicate the path to the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

### **3.3.2 Parameter: tiptable**

Explanation: Indicate in this field the type of CAPFITOGEN data table to be analysed and, if required, corrected. The following are the options:

- Passport data table (see 3.1.1)
- External source data table (see 3.1.2)
- Phenotypic data table (see 3.1.3)
- Genotypic data table (see 3.1.5)
- Type of phenotypic variables data table (see 3.1.4)

### **3.3.3 Parameter: pasaporte**

Explanation: For *local mode*, indicate in this field the name of the file with the table to be analysed (and/or corrected) in text format followed by the '.txt' extension. For example, if the name of the file is 'table' it should be written as 'table.txt'. Remember that this file should have been previously saved in the 'Pasaporte' folder which is part of the CAPFITOGEN set of folders. For *on server mode*, only the .txt file should be uploaded or selected from those that have been previously uploaded and that are located in the User's Files and Results area.

### **3.3.4 Parameter: access**

Explanation: Indicate in this field the number of entries or accessions contained in the table to be analysed. This field is not to be considered when analysing 'phenotypic variables data tables'.

### **3.3.5 Parameter: geoqual**

Explanation: Select this option only if the passport data has been previously analysed with GEOQUAL tool and, therefore, it contains the following additional columns: COORQUAL, LOCALQUAL, SUITQUAL, TOTALQUAL, and TOTALQUAL100. This parameter applies only to 'Passport data' tables.

### **3.3.6 Parameter: availab**

Explanation: Select this option only when the table to be analysed contains the additional column 'AVAILAB' (not part of the standard CAPFITOGEN passport table). This additional column indicates the possibility for each entry to be selected to include sub-collections (ColNucleo and FIGS\_R tools). This option applies only to 'Passport data' tables.

### **3.3.7 Parameter: *fixthem***

Explanation: Select this option to automatically correct the errors detected. If selected, a new table with all the errors corrected and ready to be used with other tools will be saved in the 'Pasaporte' folder. A report showing the problems encountered will also be generated in the same folder.

### **3.3.8 Parameter: *nmark***

Explanation: Fill in this field if the table to be analysed is either a phenotypic or a genotypic data table. Indicate here the number of phenotypic or genotypic variables that the table to be analysed must have.

### **3.3.9 Parameter: *phenot***

Explanation: Applies only to 'phenotypic variables data tables'. Indicate here the name of the table with the phenotypic information you wish to analyse. Remember that all tables, including this one, should be saved within the 'Pasaporte' folder. As for the parameter in 3.3.3, do not forget to include the file extension, in this case, '.txt'.

## **3.4. Results of TesTable**

One or two tables will be found in the 'Pasaporte' folder (*local mode*) or in the User's Files and Results area (*on server mode*) depending on whether the errors in the table were chosen to be automatically corrected or not.

### **3.4.1 'TestTableAnalysis.txt':**

This table will always be generated, even if the user selects the automatic correction or not. It contains a report of the possible errors that the analysed table can contain. The report is divided into different sections, all with information in both Spanish and English (Fig. 16). In a 'passport' table, the sections are the following:

**3.4.1.1 Dimension:** The number of columns is checked according to the type of table and its characteristics (e.g., if it contains the fields GEOQUAL and AVAILAB).

**3.4.1.2 Column headings:** Checks that column names or headings are correct (especially those of a 'passport' table).

**3.4.1.3 Ghost rows:** Checks there are no empty rows at the end of the table. These 'ghost rows' normally appear when the original table in Excel has empty rows, for example as a result of having previously deleted rows. If the Excel file containing empty rows is then saved as a .txt file, the ghost rows will consequently appear generating errors when using any of the tools.

**3.4.1.4 Blank cells:** As mentioned above, blank cells are not allowed in any of the tables used by CAPFITOGEN tools. In this section, both the existence of blank cells and their location are reported.

**3.4.1.5 Information required in each field:** Here, it is verified that the data entered into each field follow the required standards to be analysed by CAPFITOGEN tools. For instance, if the field requires a date (e.g., COLLDATE) and the data entered does not meet the required format, it will be reported in this section. The fields to be analysed are:

- ACCENUMB
- ORIGCTY
- ADM fields (locality descriptions through administrative divisions)
- Coordinates in decimal (DECLATITUDE/DECLONGITUDE) and sexagesimal (LATITUDE/LONGITUDE) format
- GEOREFMETH
- COLLDATE
- SAMPSTAT
- COLLSRC
- AVAILAB

Similarly, the 'TestTableAnalysis.txt' report for other types of tables will show different sections for each important field that the target table contains.

### **3.4.2 'Pasaporte Passport corr.txt':**

This table will be generated if after having inserted a 'Passport' table, the option 'fixthem' is selected. It corresponds to the passport table automatically corrected by TesTable. This table will be suitable to be used in the several CAPFITOGEN tools that require this type of table (all except ELCmapas).

### **3.4.3 'FuentesExt ExtSources corr.txt':**

This table will be generated if after having inserted a 'FuentesExternas' ('external source') data table, the option 'fixthem' is selected. It corresponds to the species occurrence external source data table automatically corrected by TesTable. This table will be suitable to be used in the Representa tool (Chapter 8).

### **3.4.4 'Genotipo Genotype corr.txt', 'Fenotipo Phenotype corr.txt' or 'NaturalezaVariables Variable-Type corr.txt':**

This table will be generated if the option 'fixthem' has been chosen after having inserted a 'phenotypic', 'genotypic' or 'type of phenotypic variables' data table. It corresponds to either the phenotypic or genotypic table automatically corrected by TesTable. The table will be suitable to be used in DIVmapas tool (Chapter 9). These phenotypic and/or genotypic characterization tables will be required by DIVmapas depending on the kind of diversity map to be created.

## 3.5. Final considerations

It is important to highlight that obtaining a corrected table by choosing the option 'fixthem' in TestTable does not mean in any case that TesTable will improve the quality of the data. This tool will only delete formatting errors and replace them with the text 'NA', except in the field ACCENUMB in which 'NA' will generate errors itself. Errors in the field ACCENUMB can only occur due to the existence of duplicates or 'NA' values. In this case, TesTable will generate a new value (numeric or alphanumeric) by adding to the duplicate or 'NA' value found the row number where the error is located; hopefully, this new value will not become a duplicate itself.



**Regional Workshop, Florianópolis (Brazil), May 2014.**





# 4 | GEOQUAL Tool

## 4.1. What is the Evaluation of the Quality of Geo-referencing in passport data?

This methodology determines the degree of certainty contained in some passport descriptors whose function is to unequivocally define the location where the germplasm was collected. GEOQUAL is thus able to assess the quality of the data describing the location and the coordinates indicated as a collection site.

In broad terms, the concept of quality applied to data has received different definitions. In the geographical context, the definition of quality as 'fitness for use' or potential for use is widely accepted (Chrisman, 1983). This relates quality to the possibility of using data. The uncertainty associated with all kinds of data is a property of anyone who obtains or uses it rather than of the data itself. Therefore, quality and uncertainty share a degree of variable subjectivity, which can be reduced to a certain extent by using methodologies that perform evaluations on as objective a basis as possible. In any case, quality and uncertainty are taken as measures of understood risk and assumed risk (Chapman, 2005).

The need to assess the quality of the geo-referencing of information available about the presence or absence of biological entities is a tangible issue in a range of different areas such as ecology, spatial analysis, and the patterns of the distribution of species. Many studies point out that quality is a critical issue in methodologies such as the modelling of the distribution of species. The certainty of the occurrence of a species at a given site is crucial for any method using presence or absence as raw data (Foley *et al.*, 2009; Hill *et al.*, 2009; Otegui *et al.*, 2013).

An estimate of the degree of uncertainty in the geo-referencing of sites concerning the presence or absence of species then becomes a key aspect before any analysis that uses spatial aspects to study distribution. Many analyses of this kind lead to decision-making about the practical aspects of areas such as the conservation of biodiversity. Therefore, the introduction of reliable baseline information to feed into the appropriate analysis will produce reliable results as well as successful and timely decisions.

## 4.2. History of GEOQUAL tool

The methodology that gave rise to GEOQUAL is the result of four years of development, from the moment when the need arose for an estimator able to measure the reliability (or risk, whichever fits) of the geo-referencing of a collection site, usually reflected in passport data. This need arose at the end of 2009 when the passport data for the Spanish National Inventory of Plant Genetic Resources were being prepared to be ecogeographically characterized. At the time, obtaining an idea of the quality of the geo-referencing for passport data was a priority for the creation of the System for Ecogeographic Information for Spanish Plant Genetic Resources (Sistema de Información Ecogeográfica de los Recursos Fitogenéticos - SIERFE).

SIERFE was a system that enabled the selection of germplasm based on the environmental characterization of a collection site through an internet portal. With the development of GEOQUAL and its incorporation into SIERFE, a qual-

ity estimator allowed SIERFE users (seekers of germplasm, such as breeders, scientists, or farmers) to define their requirements in terms of the quality of georeferencing when selecting germplasm by ecogeographic variables. At the time, this represented a major advance in the development of information systems and germplasm selection. Over 45,000 accessions in the Spanish inventory were ecogeographically characterized and each one was given a quality rating value on a scale of 0 to 100.

GEOQUAL was then specifically tailored to the characteristics of the passport data from the Spanish National Inventory of Plant Genetic Resources passport. This was possible using a range of programs, most of which are commercial-type programs such as ESRI's ArcGIS.

In 2011, within the framework of the PGR secure project enshrined in the Seventh Framework Program of the European Union (<http://www.pgrsecure.org>), it was necessary to clear four databases containing information of the occurrence of wild varieties and species related to four taxa of agricultural interest in Europe (*Avena*, *Beta*, *Brassica*, and *Medicago*). More than 33,000 records received a GEOQUAL value, which meant that the quality of some 4,000 accessions could neither be considered nor improved. Since then, several European researchers on agrobiodiversity issues became interested in GEOQUAL, which resulted in a demand for the development of a user-friendly management tool permitting the application of GEOQUAL to different formats of species presence data.

In 2012, when the development of CAPFITOGEN tools began, GEOQUAL was selected as part of the priority applications. The idea addressed the challenge of creating a tool capable of evaluating the quality of geo-referencing data; a simple tool that already had all the necessary information preloaded and that did not require a great knowledge of geographic information systems (GIS) to apply it. Additionally, the tool had to offer an integrated solution (using only a GIS program) and employ the passport descriptors format defined by the FAO and Bioversity International in 2012 as a basis. Finally, it had to be capable of being transferred to technicians of national programs. The GEOQUAL tool presented here is the evolution of an original idea transformed into easily adopted technology that offers a range of adaptability factors that are appropriate for the conditions and needs of various national programs for the conservation of plant genetic resources.

### 4.3. Features of GEOQUAL

GEOQUAL tool comprises four parameters, three of which provide different approaches to the quality of georeferencing (COORQUAL, SUITQUAL, and LOCALQUAL) and a fourth parameter (TOTALQUAL) that summarizes the first three. The base parameters are calculated in ranges from zero to twenty, with zero being no quality and 20 maximum quality. Sometimes, depending on the passport data available, the calculation of LOCALQUAL can be sidestepped, as explained later. Additionally, the program has generated a parameter transforming TOTALQUAL's initial values (0 to 40 or 0 to 60) into an evaluation range from 0 to 100, to make it easier to use and interpret the evaluation values (TOTALQUAL100). It is important to note that GEOQUAL operates with the FAO-Bioversity passport descriptors format published in 2012 with the addition of four location descriptors (ADM1, ADM2, ADM3, and ADM4) that correspond to different administrative figures by country. However, if the data were in the 2001 FAO-IPGRI format, GEOQUAL would also be able to operate after migrating the 2001 formatting information to 2012 without having to

add information for new fields included by the 2012 version. Nevertheless, it would have to consider including the four ADM descriptors.

GEOQUAL tool includes a model of a table of passport descriptors based on the FAO-Bioversity's multi-crop descriptors with the addition of the four ADM descriptors in Excel format. This model can be found in the folder 'Formatos\_Formats', subfolders 'Español' and 'Formatos GEOQUAL', file 'Tabla\_pasaporte\_modelo\_FAO\_Bioversity 2012 modificada.xls', where colour green is used to identify descriptors that are essential for GEOQUAL and yellow is used to identify those that, although not essential, are nonetheless important. Non-designated (colourless) fields are not considered by GEOQUAL but their position in the table should be maintained (as in the case of those listed) so that GEOQUAL can find the variables it needs to perform the analysis exactly where it expects to find them. As a general rule, when filling out this table, when it is not clear what information is being sought, it is best to write NA in the requisite field, which normally means Not Applicable, but in the case of GEOQUAL also indicates that there is no information available.

### **4.3.1 Description of GEOQUAL's base parameters**

#### **4.3.1.1 Parameter COORQUAL**

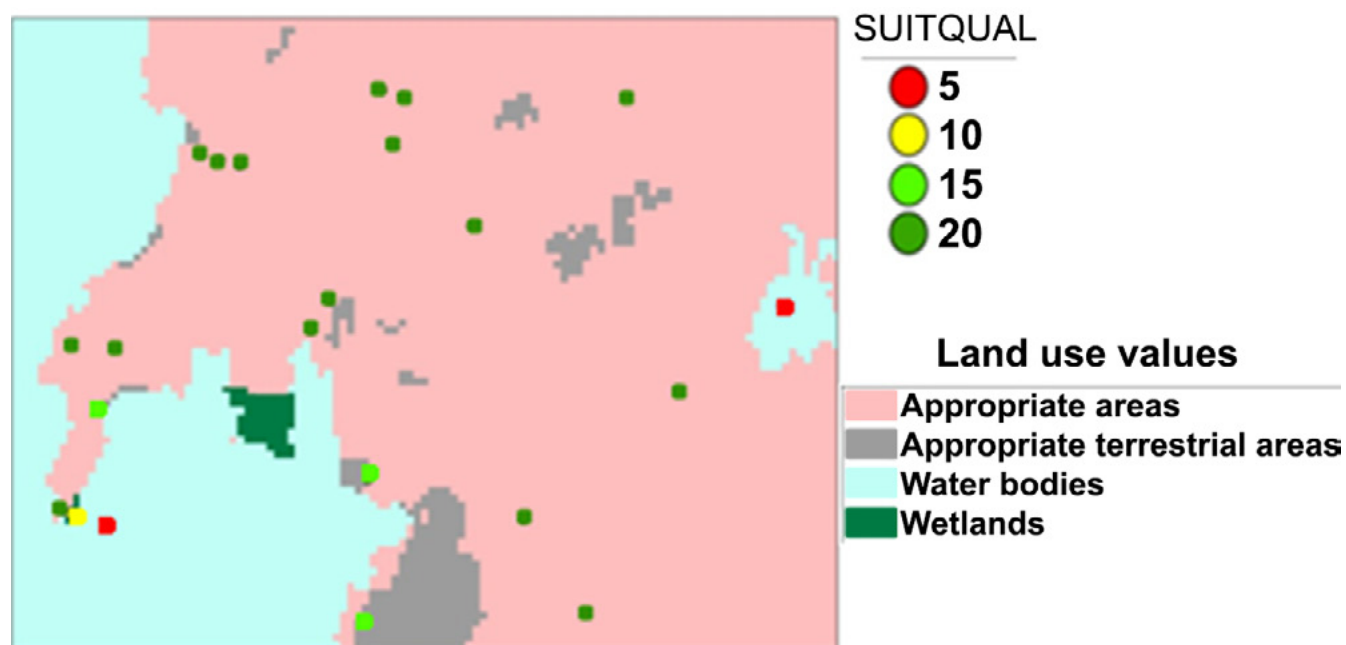
This parameter determines the intrinsic quality of the coordinates contained in the passport data. Four sub-parameters are initially used to determine it:

- a) ERRORES:** If the coordinates in decimal or sexagesimal format contain values out of the references of the WGS84 lat-long coordinates system. It uses the descriptors LATITUDE, LONGITUDE, DECLATITUDE, and DECLONGITUDE.
- b) PRECIS:** This applies to coordinates in a sexagesimal format that comply with the coding of the list of FAO-Bioversity 2012 passport descriptors. This sub-parameter determines whether the coordinates were obtained with an accuracy of seconds, minutes, or degrees.
- c) GEORBLE:** The descriptor evaluates the possibility of obtaining coordinates of the collection site from the available data describing the location.
- d) INTERTEMP:** It uses the COLLDATE descriptor values and interprets them according to the possibility of using geo-referencing methods. For example, for collections that occurred after 2000, it is highly likely that GPS was used, which would increase the quality of the coordinates.
- e) GEOREFMETH:** It assesses the system used to assign coordinates to the collection site. GEOREFMETH corresponds to a field in the FAO/Bioversity 2012 passport table. This sub-parameter will only be considered when there are values available for all accessions in this field.

Each sub-parameter provides an evaluation on a scale of zero to three, where zero corresponds to minimum quality and three to maximum quality. The combined values of each sub-parameter generate the parameter COORQUAL in a range from zero to twenty.

### 4.3.1.2 Parameter SUITQUAL

This parameter assigns a quality value to coordinates according to how appropriate the collection site is for plant growth. It differentiates the nature of the accession (wild or cultivated according to the SAMPSTAT descriptor). Information about the characteristics of the collection site comes from a land use map (Global Land Cover 2000 or GLC2000). This is the oldest freely accessible global coverage map on land use with an appropriate resolution (1 km). The original classes of this map change according to how appropriate each class is for the presence of cultivated or wild plants, on a scale of 0 to 20.

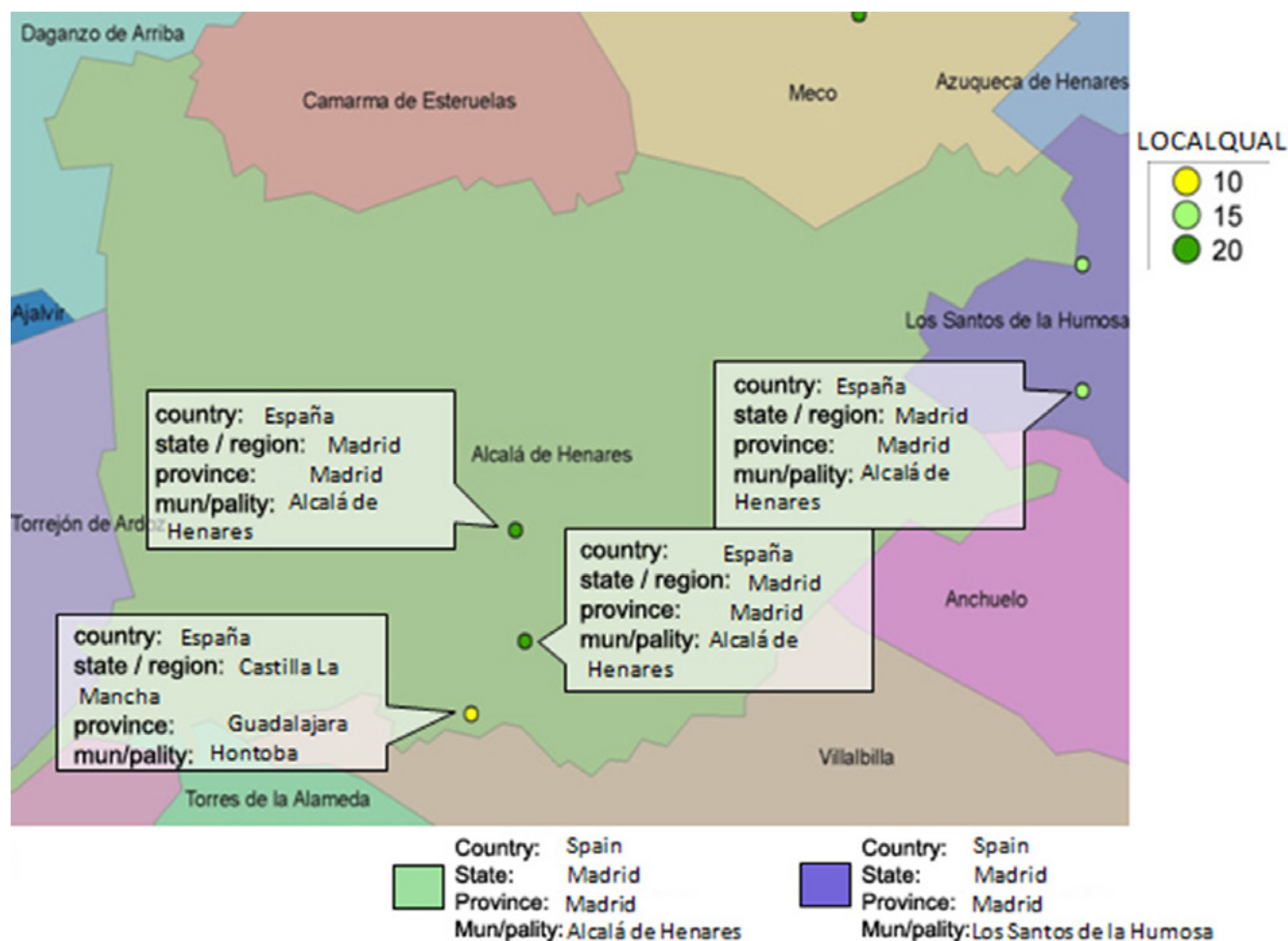


**Figure 22.** An example of how to obtain SUITQUAL values according to the interpretation of land use values.

### 4.3.1.3 Parameter LOCALQUAL

LOCALQUAL is the result of the comparison between the location description where germplasm was collected from the fields ORIGCTY, ADM1, ADM2, ADM3, ADM4, and COLLSITE, with fields ISO, NAME1, NAME2, NAME3, and NAME4 from the database called 'Global Administrative Areas' (GADM) v2.0. These were drawn using the coordinates provided by DECLATITUDE and DECLONGITUDE (or through the transformation to the decimal format used by LATITUDE and LONGITUDE). Unlike the process that 'Check Coordinates' (checking coordinates) performs, included in DIVAGIS, where the comparison is absolute (the terms must match character by character to be considered a match), GEOQUAL uses the generalized Levenshtein distance through the 'agrep' function of the base package of R, which takes into account the number of insertions, deletions, or changes of characters between the two strings being compared. Thus, even allowing for a certain number of such changes, the 'agrep' function can identify concordances despite typographical errors or differences created by using alphabetical characters from certain languages that are not encoded properly (such as the 'ñ' and the accents in the Spanish language).

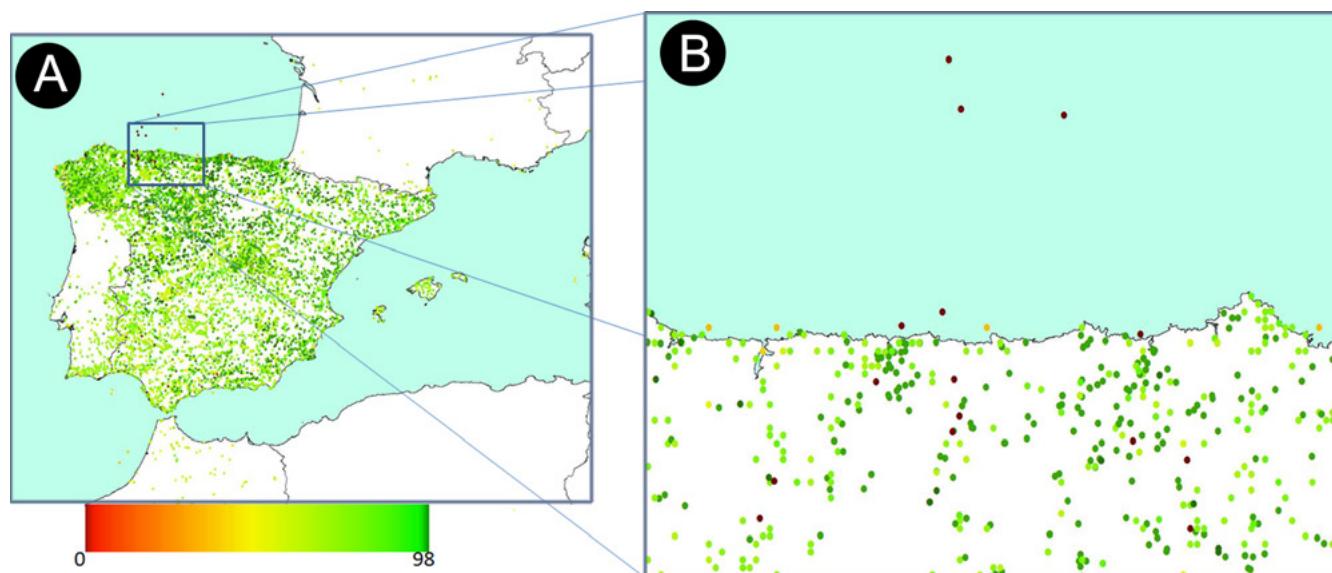
To be on the safe side, LOCALQUAL also compares the fields included in GADM (VARNAME1, VARNAME2, VARNAME3, and VARNAME4), which are variants of the official name of the administrative unit and may be used by curators when registering germplasm in their passport databases. Lastly, LOCALQUAL considers the series of positive comparisons between different pairings (ORIGCTY with ISO, ADM1 with NAME1, etc.) to calculate a value on a scale of zero to twenty.



**Figure 23.** An example of obtaining LOCALQUAL values according to the comparison of administrative levels of the data provided by the user and the coordinates drawn from GADM.

#### **4.3.2 Description of parameters TOTALQUAL and TOTALQUAL100**

The final summary parameter of TOTALQUAL is simply the sum of the values of COORDQUAL, SUITQUAL, and LOCALQUAL. For the possible ranges of values for these three parameters, TOTALQUAL can work with values from 0 to 60. However, to make it easier to interpret and analyse the results generated by GEOQUAL, parameter TOTALQUAL100 is also calculated. This is a transformation of TOTALQUAL to a range of values from 0 to 100, where 0 is zero quality (including the lack of coordinates) and 100 represents an optimal theoretical quality.



**Figure 24.** Results of the application of GEOQUAL to the Spanish National Inventory of Plant Genetic Resources. The values reached by TOTALQUAL100 are displayed. A) About 45,000 accessions from the inventory evaluated by GEOQUAL, with TOTALQUAL100 values between 0 and 98. B) Approach to the coastal area where it can be observed how the TOTALQUAL value decreases as the point moves into the sea.

### **4.3.3 Determination of quality thresholds**

Since GEOQUAL was first used, it has been designed to be applied as a highly objective methodology, where the user has only a minimum intervention in the achievement of the final value. However, any determination of quality involves subjective components and GEOQUAL is no exception.

For example, there is a degree of subjectivity when certain values concerning the suitability of growing plants are applied to certain categories of land use. Also, the definition of the point from which values may be considered to be high or low is a subjective matter that has to do with the observer rather than the technique. The threshold over which an accession is considered to be correctly georeferenced using GEOQUAL values must be defined by the user of the data, based on their expectations and needs. Different thresholds may be set, depending on how the data are to be used, how they will be studied, and the degree of accuracy and precision of the information provided by the other sources. It is advisable to see how the TOTALQUAL100 values are distributed in the set of accessions as a whole, to know in advance that an over-demanding threshold (near 100) will result in a small selection of accessions, whereas one that is less demanding (under 50) will lead to a larger selection of accessions.

## 4.4. Using GEOQUAL tool

Once the parameter scripts and GEOQUAL tool have been loaded in RStudio or said tool has been selected in *on server mode*, it will be necessary to define a set of parameters to ensure the R program runs correctly.

After all the parameters and paths required by GEOQUAL are defined, the tool analysis process will start after clicking on the 'Run' button (*local mode*) or 'Start' (*on server mode*).

After some time that may vary due to the introduction of specific resolution parameters, the type of analysis, the amount of processed data, or the computer's hardware settings, GEOQUAL will produce results to be stored where indicated (parameter 4.4.1.5 for *local mode* or in the User's Files and Results area in *on server mode*).

### 4.4.1 Initial parameters defined by the user

#### 4.4.1.1 Parameter: ruta (only for local mode)

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### 4.4.1.2 Parameter: pasaporte

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is called 'table', you should enter: 'table.txt'. Remember to save the file first in the 'Pasaporte' folder which is part of the set of folders making up the CAPFITOGEN directory. For *on server mode*, you should only upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

#### 4.4.1.3 Parameter: precision

Explanation: Select high- or low-resolution maps to determine whether the coordinates for a collection site fall in the sea and if so, how far in. High resolution may slow the process down a little when working in very large databases (over 15,000 accessions with coordinates).

#### 4.4.1.4 Parameter: local

Explanation: Indicate whether you wish to use parameter LOCALQUAL to evaluate the quality of the geo-referencing. LOCALQUAL is a parameter of comparison between a locality described and drawn by GIS. If your data does not contain any description of locality, or if the description is completely contained in the field COLLSITE, this option is UNSUITABLE.



#### 4.4.1.5 Parameter: resultados (only for local mode)

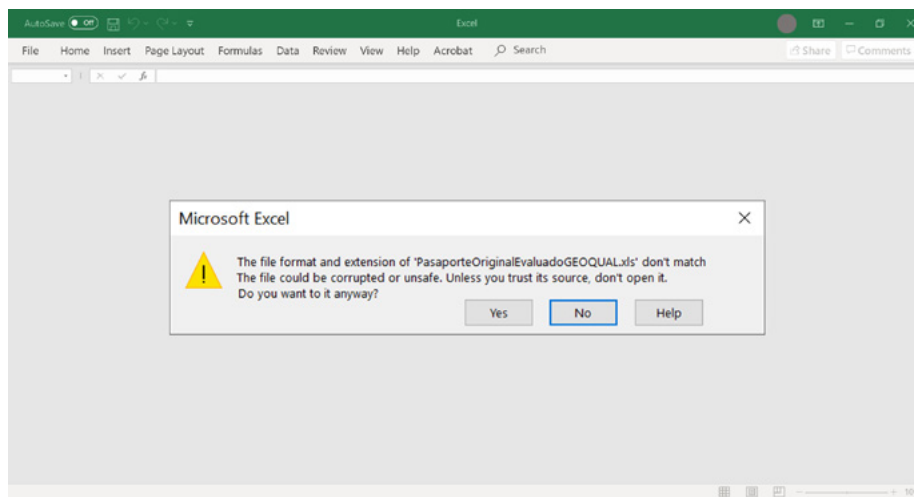
Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 4.5. Results of GEOQUAL

In the path and folder created for 'resultados' (parameter 4.4.1.5 in *local mode* or in the User's Files and Results area in *on server mode*, there should be three tables, a map of vector-type points (shapefile) and a file with the list of parameters used in the execution of GEOQUAL ('Parameters\_Geoqual.txt'). In *on server mode*, an execution progress report is also included in a file called 'Error\_process\_info.txt'.

### 4.5.1 Tables

The tables generated by GEOQUAL are in tab-delimited text format with two extensions (.txt and .xls). The operating system will try to open the .txt files with the text editing software associated with this extension. In the case of a .xls file, if the user has MS Excel<sup>®</sup> installed, this software will try to open it, but since it is a text file, a warning message will appear (see Fig. 25). This message can be ignored, and the table can finally be displayed.



**Figure 25.** MS Excel<sup>®</sup> warning message about the .xls file (which is a tab-delimited text file) to be opened. At this point, the user must click on the option 'N' and the table will open correctly

**4.5.1.1 'PasaporteOriginalEvaluadoGEOQUAL.txt' and .xls:** It is the passport table in the suggested format, that was originally used for analysis, with the addition of five columns with the values obtained for parameters SUITQUAL, LOCALQUAL, COORQUAL, TOTALQUAL, and TOTALQUAL100.

**4.5.1.2 'tabla\_de\_analisisGEOQUAL.txt' and .xls:** This table also contains all columns of the passport table that were originally introduced for analysis purposes, although in this case only those accessions with coordinates are included. However, the most important aspect of this table is that it includes all the columns that correspond to extractions, interpretations, or sub-parameters and that are considered necessary to calculate the values of GEOQUAL parameters. The list of additional variables included in this table and their explanation are found in Annex 18.4.

**4.5.1.3 'StatsSummaryGEOQUAL.txt' and .xls:** It is the table that shows the averages, mode, maximum and minimum values, and standard deviation for SUITQUAL, LOCALQUAL, COORQUAL, TOTALQUAL, and TOTALQUAL100 for the evaluated passport data.

## **4.5.2 Maps**

**4.5.2.1 Point map in a vector format of the 'shapefile' type.** This map is accompanied by a table that includes the values of the GEOQUAL evaluation parameters in such a way that the points can be shown in different colours according to their score (quality) when using DIVA-GIS. A 'shapefile' is made up of up to 6 files of the same name but with a different extension. In the case of GEOQUAL, the shapefile comprises just three extensions (.shp, .shx, and .dbf) and is called ShapefilePuntosGEOQUAL.

**4.5.2.2 Point map in Google Earth format.** This map corresponds to the file mapa\_puntos\_google.kml. If you have the Google Earth program installed on your computer, just double-click on its name in Windows Explorer, and a point map (in the form of tacks or pins) will open in that program, locating the collection sites on satellite images. Clicking on the thumbtacks opens a small window showing the TOTALQUAL100 value of each accession.

## **4.6. References**

Chapman, A.D. 2005. Principles of data quality, version 1.0. Report of the Global Biodiversity Information Facility, Copenhagen.

Chrisman, N.R. 1983. The role of quality information in the long-term functioning of a GIS. Proceedings of AUTOCART06, 2: 303-321. Falls Church, VA: ASPRS.

FAO, IPGRI. 2001. Lista de descriptores de pasaporte para cultivos múltiples desarrollada por la FAO y el IPGRI .

FAO, BIOVERSITY. 2015. FAO/Bioversity multi-crop Passport descriptors V.2. URL: <https://bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/>

Foley, D.H., Wilkerson, R.C., Rueda, L.M. 2009. Importance of the “what,” “when,” and “where” of mosquito collection events. J Med Entomol. 2009 Jul;46(4):717-22.

Hill, A.W., Guralnick, R., Flemons, P., Beaman, R., Wieczorek, J., Ranipeta, A., Chavan, V., Remsen, D. 2009. Location, location, location: utilizing pipelines and services to more effectively georeference the world's biodiversity data. *BMC Bioinformatics* 10 Suppl 14: S3. doi: 10.1186/1471-2105-10-S14-S3

Otegui, J., Ariño, A.H., Encinas, M.A., Pando, F. 2013. Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS One*. 8(1), e55144. doi: 10.1371/journal.pone.0055144.

Soberón, J., Peterson, T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. *Phil. Trans. R. Soc. Lond. B*. 359: 689-698.



National Workshop, Madrid (Spain), November 2013.

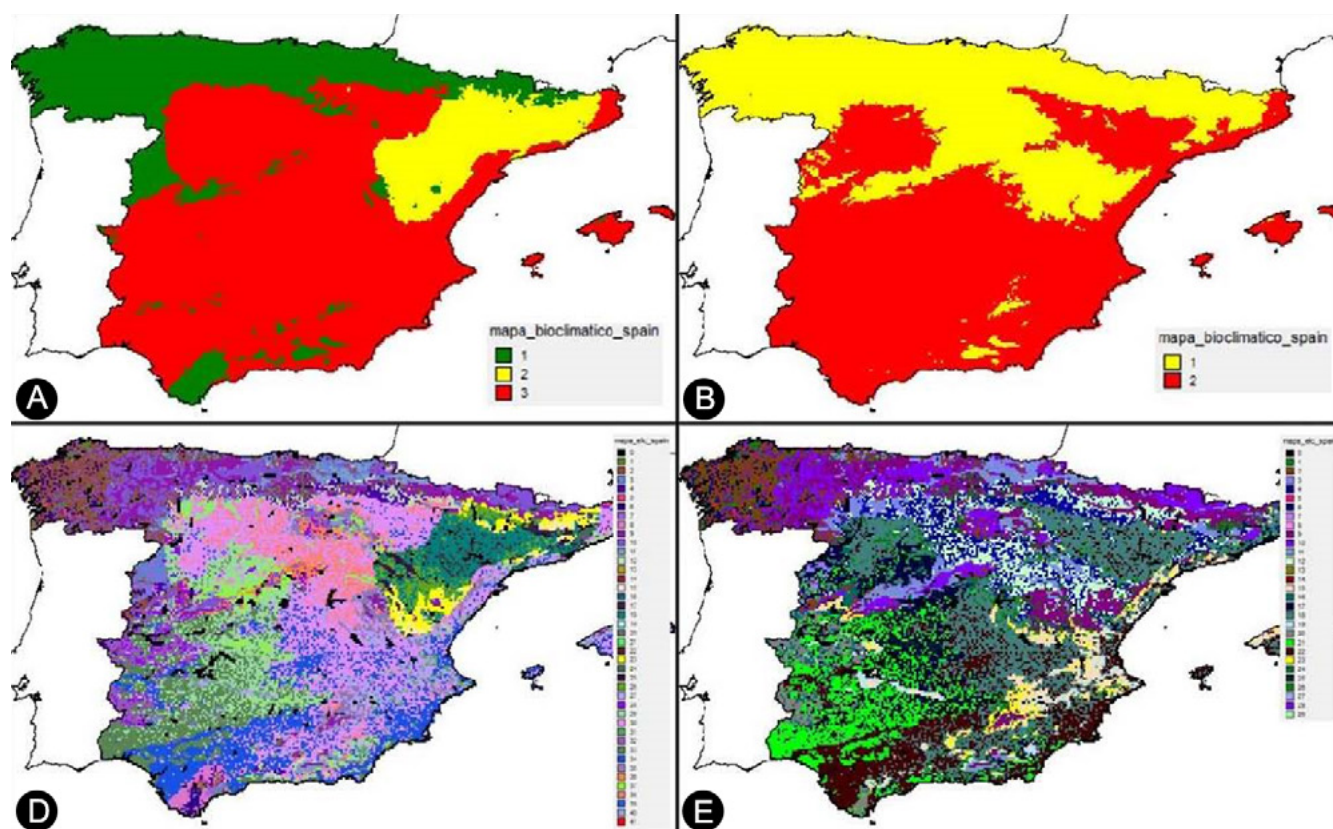


# 5 | SelecVar Tool

## 5.1. Why should we select abiotic variables with a high influence in determining adaptive scenarios for a target species?

Users of ELCmapas tool (chapter 6) have already been able to answer this question. However, it should be formulated by anyone who wants to make species distribution models such as those made by Modela tool (chapter 14). The reason for formulating this question is simple: trying to determine the different species' adaptive scenarios in a territory using all the available variables (105 in CAPFITOGEN tools) does not have a biological meaning but also slows down and alters the results of both, the multivariate analysis (performed by ELCmapas) or modelling (performed by Modela).

As an example, Fig. 20 illustrates the effects of selecting (or not) ecogeographic variables to obtain an ecogeographic land characterization (ELC) map for *Vicia sativa* in Spain. The territory will appear divided differently when using all available bioclimatic variables compared to the selection of a specific set of those variables.



**Figure 26.** ELCmapas tool (chapter 6) results showing maps with and without previous selection of variables. A) Map showing bioclimatic zones after selecting 3 variables. B) Map showing bioclimatic zones using all 67 available variables. C) ELC map with bioclimatic variable selection. D) ELC map using all available bioclimatic variables.

Selecting, or not, ecogeographic variables to be used in species distribution models can strongly affect the results of this kind of analysis (Austin & Van Niel, 2010). An interesting study shows the significant effect that using different sets of predictive abiotic variables for modelling plant species distributions can have on the results (potential distributions), spatial patterns, and predictions (Pliscoff *et al.*, 2014). Thus, selecting predictive variables can make the predicted occupied area fluctuate up to 2,000%, and the species extinction risk estimates up to 50%.

## 5.2. Selection of variables for ELC maps

There are two types of ELC maps: general and specific (see section 6.4.). For general maps, the selection of variables is a simple process. It can simply consist of asking experts for some advice on the variables that can affect plant distributions in a broad sense. However, to generate specific maps, not only expert knowledge is needed for the selection of abiotic variables, but also a careful selection of variables supported by statistics. For this reason, the final set of selected variables should meet, if possible, the requirements outlined in the sections below. Note that these requirements can also be applied to those variables used as predictive in a species distribution model.

### **5.2.1. Heterogeneity and minimum redundancy**

A cluster analysis (in the case of an ELC map) or a distribution model must be fed with the minimum number of variables that can explain the occurrence of different adaptive scenarios to which a target species or population presence/absence responds. A small number of variables allows for faster and more efficient processes and analyses, also facilitating the interpretation of the results. It is possible to discard variables that do not contribute to the analysis or variables that provide the same information as others.

Homogeneous variables are not useful to create adaptive groups or to generate potential distributions. Thus, when extracting ecogeographic information from different variables for a set of collection sites or species' populations, those variables with the same values for all sites or populations can be removed.

It is possible to determine what variables provide identical or very similar information through the use of bivariate correlation analysis (BCA), or correlation of each pair of variables of the initial set considered. This way redundancy is avoided when obtaining ELC maps or modelling distributions as only one variable is left representing them all.

A principal component analysis (PCA) also allows detecting variables that provide similar information. This type of analysis has been designed to reduce dimensionality in a set of data with a large number of variables. It uses an orthogonal transformation to convert the original variables into linearly uncorrelated variables called principal components (new variables).

PCA is used as a tool in exploratory data analysis. It gives information related to both the variability of the original data explained by each component (eigenvalues) and the weight of each original variable on each component (eigenvector).

For the selection of variables using PCA consider a small set of principal components (e.g., a set that represents more than 70% of the original variance) and determine the original variables with the highest influence in absolute values (i.e., without considering any positive or negative directions). Amongst a few principal components, the selection of one or two original variables with the highest loading in its principal component will give a reduced set of variables that are highly discriminatory and uncorrelated (similar to method B4 described by King & Jackson, 1999).

### **5.2.2 Contribution of variables**

The aim is to determine the contribution made by each variable considered when defining adaptive groups or the species distribution.

When defining the potential adaptive scenarios for a target species in a territory, it is very useful to identify first the abiotic variables that can influence the most the clustering of populations that have been either collected or identified in areas with different adaptive features. Different approaches can be used here. SelecVar uses the Random forest technique (classification).

Random Forest (RF) is a powerful statistical classifier. It provides very accurate classifications (clusters) in environments with complex interactions amongst variables. It also allows determining the importance of the variables included in the classification (Cutler *et al.*, 2007). The technique to determine variable importance has been implemented in the function ‘importance’ of the R-package ‘randomForest’. A table is generated as a result of using the function. The table contains as many rows as the number of variables analysed, and only two columns showing the measures of importance, such as mean decrease accuracy or mean decrease Gini. According to Cutler *et al.* (2007), the mean decrease accuracy for a variable is the normalized difference of the classification accuracy for the out-of-bag data when the data for the variable is included, and when the values of the variable have been randomly permuted. Therefore, higher values of mean decrease accuracy indicate variables that are more important to the classification. In the study of the importance of climatic variables selection to obtain reliable predictive species distribution models, RF was used to determine sets of six variables with the highest values of mean decrease accuracy.

## **5.3. What does SelecVar do and how to use its analyses?**

SelecVar is designed to offer users enough objective arguments to carry out a selection of variables that is useful for generating ELC maps (see chapters for ‘ELCmapas’ or ‘Tzones’ tools) or species distribution models (see chapter for ‘Modela’ tool). It performs, at the user’s request, the selection of variables using Random Forest (RF), bivariate correlation analysis (BCA), or principal component analysis (PCA). The first two analyses are run sequentially (first RF and then BCA) and independently for each ecogeographic component (bioclimatic, geophysical, and edaphic variables).

The selection of variables for either ELC maps or species distribution models can ideally be an objective-subjective process. SelecVar would be the objective side, and the subjective one would be expert knowledge (i.e., a survey on



the relevance of the variables amongst experts in adaptive aspects of the target species). Literature compilations on adaptive aspects of the species can also be useful within the subjective process.

The use of SelecVar to determine the final list of variables to be included in ELCmapas or Modela may vary according to the user's needs or point of view on the analyses carried out. For CAPFITOGEN2, SelecVar allowed the user to perform all possible combinations of analysis among PCA, BCA, RF, and another procedure called *clusvarsel*. However, this last procedure was discarded in CAPFITOGEN3 (both modes of use) due to the high demand for computational resources for its execution and the many errors that occurred. Thus, SelecVar was modified in such a way that in version 3 it always performs the following steps:

1. RF: A fraction of the total variables that the user initially indicates the tool to evaluate is selected using the mean decrease accuracy index. The user indicates the size of the fraction of variables to be selected. SelecVar lists the variables from the most to the least important and selects the most important user-specified fraction of variables.
2. BCA: The set of variables selected by RF is subjected to a bivariate correlation test. The process of discarding variables by correlation (redundancy) begins by comparing the selected variable of least importance according to RF with the rest of the variables of greater importance. Having a correlation coefficient value and a P value (also provided by the user) as references, the variable is discarded if it is correlated with at least one of the most important ones. If the variable does not correlate with any other, it is selected. The process continues with the next variable of less importance according to RF, which is compared with the others of greater importance and discarded in case of correlation with any other. Thus, the important variables by RF continue to be discarded or selected until the process stops when the fraction (indicated by the user) of variables to be selected by bivariate correlations is reached.
3. PCA: Finally, all the variables that the user has wanted to include in the analysis are subjected to principal component analysis. In this analysis, the tool does not discard variables, but rather offers the user information on eigenvectors and eigenvalues. It also provides a table where the position of each variable selected by RF-BCA is indicated regarding the eigenvectors for the principal components the user wants.

## 5.4. Using SelecVar tool

Once the parameter scripts and SelecVar tool have been loaded in RStudio or said tool has been selected in *on server mode*, the user needs to define a series of parameters for the R programming to work correctly.

After defining all the parameters and paths that SelecVar requires, the tool will start its analysis process when you click on the 'Run' (*local mode*) or 'Start' (*on server mode*) buttons.

SelecVar will produce results after some time that may vary due to the introduction of some resolution, type of analysis, size of processed data, or computer hardware configuration parameters. SelecVar will save these results where indicated (parameter 'resultados' for *local mode* or in the User's Files and Results area in *on server mode*).

## **5.4.1 Initial parameters defined by the user**

### **5.4.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools are to be found. Note: use / (slash) instead of \ (backslash) to indicate the path to the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

### **5.4.1.2 Parameter: pais**

Explanation: Select the country/region where all or most of the data accessions you wish to analyse were collected. If accessions have been collected from more than one country, you may select a region, subcontinent, or continent (these options will be added progressively). You can also use rLayer tool to produce your own work frames and select them here.

### **5.4.1.3 Parameter: pasaporte**

Explanation: For *local mode*, type the name of the file containing the passport table in text format without forgetting to include the file extension (.txt). For example, if the file is named 'table', you should write 'table.txt'. Please remember that this file must first be saved in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders, which must be reflected in this field. For example, if your table is called 'table.txt' and is located in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in the field, always using / (slash) before \ (backslash) in the path description. For *on server mode*, the user only has to upload the .txt file or select it from the files that have been previously uploaded and that are located in the User's Files and Results area.

### **5.4.1.4 Parameter: disdup**

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites represent the same population. The value zero (by default) excludes accessions with identical coordinates from the representativeness analysis. The determination of the distance depends on biological (gene flow) and spatial (mean population sizes) conditions. This is a specific parameter for the target species, and it will often be necessary to consult an expert for his/her concept.

### **5.4.1.5 Parameter: geoqual**

Explanation: Select this option if the passport data have been analysed using GEOQUAL tool and thus contain 50/51 columns (rather than the 45/46 in the basic passport model). To select this option please use the table produced by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name) as a passport table. Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### 5.4.1.6 Parameter: totalqual

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

#### 5.4.1.7 Parameter: buffy

Explanation: Mark this option if you wish ecogeographic information to be extracted from a circular area around the collection site. Leaving this option unchecked means that information is only extracted from the point indicated by the collection site coordinates. This option is very useful, for example, when most of the collection sites correspond to markets and not directly to the growers' farms, or when the actual location of the collection sites is uncertain even though GEOQUAL can grant high georeferencing quality. This feature is explained in the chapter dedicated to ECOGEO tool.

#### 5.4.1.8 Parameter: tamp

Explanation: Applies only if 'buffy' has been set as TRUE (✓ in *on server mode*). Specify the radius (in km) of a circular area around the point indicated by the collection site coordinates from which the ecogeographic information is to be extracted. The values extracted from the circular area will be averaged to obtain a single value. The information will be extracted from those cells whose centroid is within the circular area.

#### 5.4.1.9 Parameter: resol1

Explanation: Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution but requires greater computing capacity than 5x5 km, although this aspect is not as limiting as in ELCmapas tool. Resolutions of 10x10 and 20x20 km may only be used for large countries, subcontinents, or continents.

#### 5.4.1.10 Parameter: bioclimv

Explanation: List (*local mode*) or select (*on server mode*) the bioclimatic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of bioclimatic variables'); copy the names and paste them separated by a semicolon (;). You will also find a blocked line with the 19 bioclim variables ready to use; simply remove the initial # symbol. In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'bioclimv'. To know the codes, names, and brief descriptions of the variables, consult the 'Variables names – Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/wp-content/uploads/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 5.4.1.11 Parameter: edaphv

Explanation: List (*local mode*) or select (*on server mode*) the edaphic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of edaphic variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'edaphv'. To know the codes, names, and brief descriptions of the variables, consult the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 5.4.1.12 Parameter: geophysv

Explanation: List (*local mode*) or select (*on server mode*) the geophysical variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of geophysical variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'geophysv'. To know the codes, names, and brief descriptions of the variables, consult the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 5.4.1.13 Parameter: latitud

Explanation: If TRUE (*local mode*) or ✓ (*on server mode*) are set, the latitude (Y) will be included as a geophysical variable to be analysed.

#### 5.4.1.14 Parameter: longitud

Explanation: If TRUE (*local mode*) or ✓ (*on server mode*) are set, the longitude (X) will be included as a geophysical variable to be analysed.

#### 5.4.1.15 Parameter: percenRF

Explanation: Enter here a value between 0 and 1. The percenRF value indicates the percentage of variables that will be selected by Random Forest (classification) and that will continue in the selection process. However, the percentage should not be expressed with a value between 0 and 100 but between 0 and 1. For example, if you want to select 60% of the variables by Random Forest so that they can continue to be selected by bivariate correlations, the value in percenRF should be 0.6. Thus, 1-percenRF will be the variables discarded as they are the least important according to the mean decrease accuracy index.

#### 5.4.1.16 Parameter: **percenCorr**

Explanation: Enter here a value between 0 and 1. The **percenCorr** value indicates the percentage of variables that will be selected by BCA in the step-by-step process described above. However, the percentage should not be expressed with a value between 0 and 100 but between 0 and 1. For example, if you want to select 30% of the important variables by Random Forest, the value in **percenCorr** must be 0.3. Thus, 1-**percenCorr** will be the variables that are discarded because they are redundant due to their correlations and less important due to Random Forest. **percenCorr** can be altered based on the minimum number of variables that the user defines in parameter 'nminvar' (5.4.1.19).

#### 5.4.1.17 Parameter: **CorrValue**

Explanation: For this parameter indicate a value between 0 and 1. This value will be used to define the threshold of what will be accepted as a high correlation according to the correlation coefficient. Thus, indicating a value of 0.5 in **CorrValue** will cause correlation values from 0.5 to 1 and -0.5 to -1 to be assumed as high correlation. Then a pair of variables with a correlation value of -0.85 would initially be redundant and that of lesser importance would be discarded. However, it must be considered that **SelecVar** will not only consider the correlation value but also its significance for discarding the variables.

#### 5.4.1.18 Parameter: **pValue**

Explanation: Enter a value of 0.001 or 0.05. **pValue** defines the significance threshold value for bivariate correlations. This value usually fluctuates between 0.001 and 0.05. The correlation will be assumed to be significant when the P value is less than that indicated in parameter 'pValue'.

#### 5.4.1.19 Parameter: **nminvar**

Explanation: Specify the minimum number of variables you want to select per component. Note that **nminvar** must be a number greater than the number of variables selected in **bioclimv**, **edaphv**, or **geophysv**.

#### 5.4.1.20 Parameter: **ecogeopcaxe**

Explanation: Since **SelecVar** performs a PCA for each ecogeographic component (**bioclimatic**, **geophysical**, and **edaphic**), **ecogeopcaxe** indicates the number of main components that are retained or the number of components for which results are offered, such as the eigenvectors. Therefore, **ecogeopcaxe** cannot be a number greater than the number of components that will be generated by the component with fewer variables. For example, if the number of main components that were generated were **bioclimatic**=5, **geophysical**=3, and **edaphic**=6, **ecogeopcaxe** must be less than or equal to three (3). Remember that the number of principal components is equal to 1-number of analysed variables.

#### 5.4.1.21 Parameter: resultados (only for local mode)

Explanation: Insert the path to the folder where the results of the analysis will be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

### 5.4.2 Results of SelecVar

After defining all the parameters and paths (for *local mode*) that SelecVar requires, click on the 'Run' (*local mode* in RStudio) or 'Start' (*on server mode*) buttons to start the analysis process of the tool.

Note that the waiting time will vary according to the type of analysis run. The results obtained by SelecVar will be automatically saved in the path and folder previously typed in 'resultados' (in *local mode*) or in the User's Files and Results area in *on server mode*. There, the user will find different folders for the bioclimatic, edaphic, and geophysical components (i.e., 'BioclimaticVariables\_pais', 'EdaphicVariables\_pais', and 'GeophysicVariables\_pais', respectively). The user will also see the 'Parametros.Parameters.SelecVar.txt' text file where the parameters used are specified, and three files with a .xls extension that correspond to text files delimited by tabulations but that can be opened directly in Excel® (Fig. 25). These files contain the finally selected variable lists. Within the folders for each component, there are new folders for each analysis performed (RF, BCA, and PCA). The same tables, lists, and result graphs will be generated for each component. The content of each analysis folder is described below.

#### 5.4.2.1 Bivariate Correlations ('BivariateCorrelations' folder) 'ConfidenceInterval\_correlation\_bioclim.xls')

requires moving the entire first row one cell to the right so that cell A1 remains empty and the table starts in cell B1.

- 'Pvalue\_correlation\_bioclim.xls': Table showing p-values for each correlation estimate. The null hypothesis will be accepted or rejected depending on the significance level chosen by the user (usually 0.05 or 0.01). In the null hypothesis, the Pearson correlation coefficient equals 0, thus there is no correlation. To properly visualize this table, move the entire first row one cell to the right so that cell A1 remains empty and the table starts in cell B1.
- 'RelevantValues\_correlation\_bioclim.txt': This table combines the results from both tables 'Estimate\_correlation\_bioclim.txt' and 'Pvalue\_correlation\_bioclim.txt' as follows: a) a value of 0 indicates that the following two conditions are met: correlation estimates between -0.5 and 0.5 (low or no correlation) and acceptance of the null hypothesis (p-values greater than or equal to 0.05), and b) original correlation values different from 0 indicate that the following two conditions are met: correlation values smaller than -0.5 or greater than 0.5 (high correlation) and rejection of the null hypothesis (P-values smaller than or equal to 0.05). The table of relevant values allows for easily checking cases of high correlation supported by statistical significance.
- 'Estimate\_correlation\_bioclim.txt': Symmetric table showing the Pearson correlation values which can range from -1 to 1. A value of 0 indicates no correlation, 1 indicates a positive correlation, and -1 negative or opposite correlation. To properly visualize this table, move the entire first row one cell to the right so that cell A1 remains empty and the table starts in cell B1.
- 'ConfidenceInterval\_correlation\_bioclim.txt': Table showing the confidence intervals for Pearson's correlation estimates for all variables. To properly visualize this table, move the entire first row one cell to the right so that cell A1 remains empty and the table starts in cell B1.

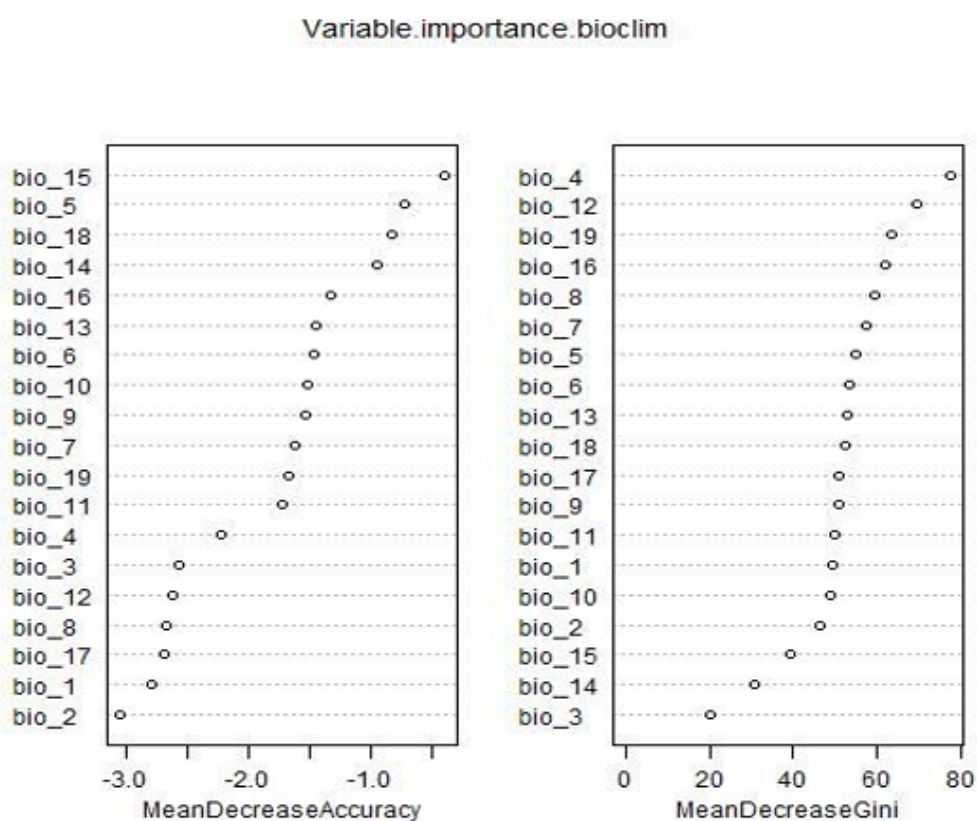
#### 5.4.2.2 Principal Component Analysis ('PCA' folder)

- 'Componente\_eigenvalues.xls': table containing the eigenvalues.
- 'Componente\_eigenvectors.xls': table containing the eigenvectors.
- 'Bioclim\_IVposition.xls': table of eigenvectors but without their original values. It contains the position occupied by the variable in each component according to its eigenvector value (absolute).

#### 5.4.2.3 Random Forest selection ('RandomForest' folder)

- 'VariableImportance\_RandomForest\_bioclim.xls': Table showing mean decrease accuracy or mean decrease Gini values. The variables can be easily sorted according to the mean decrease accuracy (from high values to low ones). The most important variables will then be at the top of the table.
- 'VarImportanceDotChart\_bio.jpeg' (Fig. 27): This chart shows the variables sorted from high (top) to low importance (bottom) according to the mean decrease accuracy or mean decrease Gini. The same kind of chart (for the mean decrease accuracy) is used in Cutler *et al.* (2007) to determine the importance of the variables.

Within the folders of each component, two files correspond to the 'accessions\_used\_componente\_varselection.xls' and 'Component\_extractedValues.xls' tables. The first table shows the accessions that were finally used to determine the importance of the variables. The second table shows the values extracted for each accession from the total of the variables that the user wanted to subject to the SelecVar analysis.



**Figure 27.** Example of the charts of mean decrease accuracy index and mean decrease Gini in the 'VarImportanceDotChart\_bio.jpeg' file.

Finally, SelecVar offers three tables: SelectedVariables\_bioclim.xls, SelectedVariables\_edaphic.xls, and SelectedVariables\_geophysic.xls. These tables show in a diagonal matrix with correlation values, and the variables finally selected by each component after applying parameters percenRF, percenCorr, CorrValue, pValue, and nminvar. These variables would be the result of SelecVar (objective selection).

## 5.5. References

Austin, M.P., Van Niel, K.P. 2010. Improving species distribution models for climate change studies: variable selection and scale. *Journal of Biogeography*, 38(1): 1-8.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. 2007. Random forests for classification in ecology. *Ecology*, 88(11): 2783-2792.

Dean, N., Raftery, A.E., Scrucca, L. 2015. Package 'clustvarsel', variable selection for Model-Based clustering. . <http://cran.r-project.org/web/packages/clustvarsel/clustvarsel.pdf>

King, J. R., Jackson, D. A. 1999. Variable selection in large environmental data sets using principal components analysis. *Environmetrics*, 10(1): 67-77.

Plissock, P., Luebert, F., Hilger, H. H., Guisan, A. 2014. Effects of alternative sets of climatic predictors on species distribution models and associated estimates of extinction risk: A test with plants in an arid environment. *Ecological Modelling*, 288: 166-177.

Raftery, A. E., Dean, N. 2006. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168-178.







National Workshop, Comayagua (Honduras), November 2013.



# 6 | ELCmapas Tool

## 6.1. What is an ecogeographic land characterization map?

In this kind of map, we can visualize various environmental scenarios that correspond to the different adaptive processes of a plant species in a given territory. The ecogeographic land characterization maps (ELC maps) are useful for the conservation and reasonable use of agrobiodiversity.

The idea of using maps to express adaptation is not new. Maps of biomes, ecosystems, and ecological regions have been in use since the middle of the last century. These maps represent environmental units comprising large and homogeneous regions. The ‘climates’ or ‘environments’ (terms used interchangeably) represented in these maps have been used to study different types of organisms (plants, animals, microorganisms). Some maps are more detailed and represent, for example, specific climates favourable for the kinds of plant formations described by Leslie Holdridge in 1947, although these were later generalized under the heading of ‘life zone classification systems’.

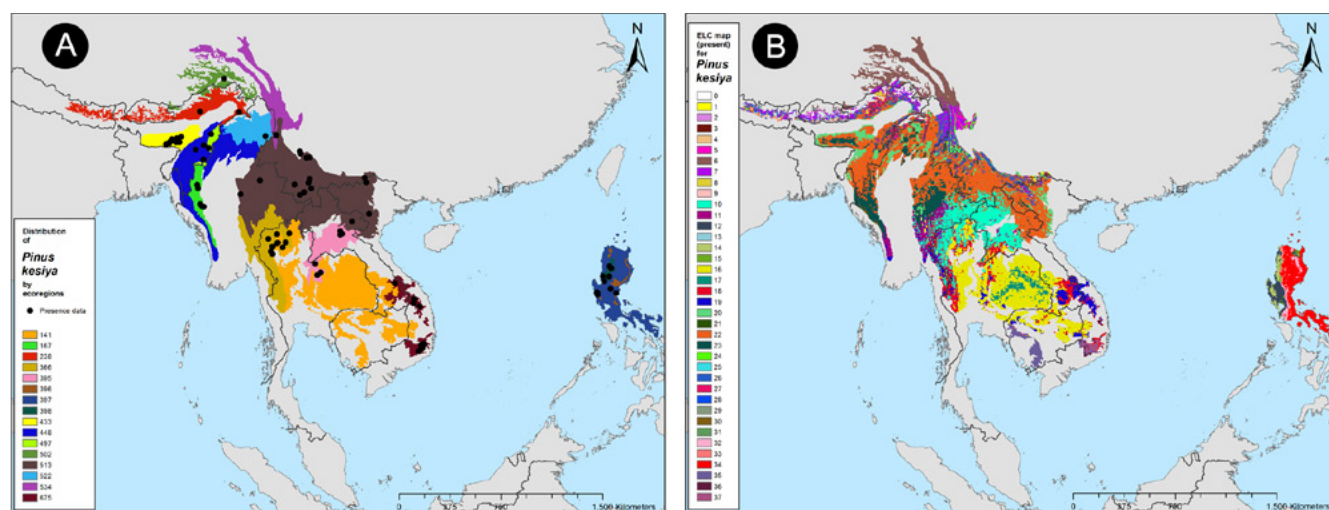
These maps have been extremely useful for biologists and naturalists studying the distribution of living organisms in relation to temperature and humidity. The Holdridge system is still employed today, for example in studies of climate change. However, the main obstacles to using this system for studies on species adaptation were its failure to differentiate between biotic features (vegetation) and abiotic ones (temperature, rainfall) on these maps, its tendency to reduce the abiotic component to only two factors, and the way in which it delimited regions (large, homogeneous, and continuous).

Designing a collection based on adaptation information or storing and using plant genetic resources according to specific efficiency criteria is nothing new, although there is little material published on this subject that explicitly refers to adaptation. One early reference is an ecogeographic map drawn up in 1997 to help create core collections (Tohme *et al.*, 1995). In this map, other different criteria were also considered after the accession selection process in addition to the ecogeographic dimension.

Since then, there have been several developments: GIS programs have become gradually more flexible and ‘user-friendly’, while some statistical packages now also include GIS utilities and tools. Similarly, the ecogeographical information available (in the form of GIS layers) is of better quality and more accessible and computer equipment with a high capacity for analysis is now sold at low prices. Furthermore, access to the internet has increased markedly in developing countries. This progress has impacted the development of maps representing different adaptation scenarios for crop wild relatives and was responsible for the generation of the first ecogeographic land characterization map for Spain in 2005 (Parra Quijano *et al.*, 2008). This is a general map that could be applied to several different crop wild relatives, although it was only used for certain species of the *Lupinus* genus. The map was obtained through multivariate analysis techniques and by determining the number of groups according to Bayesian criteria. It represented the different environmental units as small, discontinuous homogeneous regions using cross-links; these physical features were already a marked contrast to traditional bioclimatic maps. Another difference was the inclusion of geophysical and soil-type variables (in addition to the bioclimatic ones), to represent any abiotic aspects affecting plant development from the agronomic point of view.

In mid-2008 a new ecogeographic map was developed for Peninsular Spain and the Balearic Islands based on other sources of ecogeographic information, although the methodology used was similar to that produced in 2005. The

researchers wanted to ensure that the map was able to portray adaptive scenarios as faithfully as possible and establish whether it could be used to perform a validation. They evaluated the performance of the new map with eight species (four leguminous varieties and four species of grass), two of which were crop wild relatives (CWR), while the other six were local varieties. Their distribution and the “seed weight” variable as a phenotypic indicator variable with adaptive value were also used. The results were compared to two reference maps: the first displayed a physical structure similar to that of an ELC map (discontinuity, small units, and cross-links), but was created without considering aspects related to the abiotic adaptation of plants (CORINE land cover map, land use map). The second map had a different physical structure (more similar to the traditional maps) but was created for a similar purpose (DMEER map or a digital map of European ecological regions). A comparison on a more recent ecological region map can be seen in Fig. 28.



**Figure 28.** Comparison between an ecological region map and an ecogeographic land characterization (ELC) map for the *Pinus kesiya* species in Southeast Asia. A) Distribution of the known occurrences of the *Pinus kesiya* species (black points) over the 16 different occupied land ecological regions of the Global Ecoregions2017© map (Dinerstein et al., 2017). B) ELC map obtained with ELCmaps tool, of much greater detail for the species, with 38 categories.

The results of the validation study were mixed. In general, the ELC map performed better for leguminous species than for grasses, although the exception was *Zea mays* for which the map elicited a quite acceptable result. As expected, the map displayed adaptive scenarios for the two CWR but also produced satisfactory results for species composed of only local varieties, such as *Phaseolus vulgaris*. In conclusion, ELC maps provide a satisfactory rendition of adaptive scenarios and can thus be used for many activities related to the collection, conservation, and efficient utilization of plant genetic resources. However, specific maps should be created for each species or group of phylogenetically related species. Creating general-type ELC maps is not recommended when drawing conclusions from a large group of species, particularly if the map is not properly validated. It is also important to make a proper selection of the ecogeographic variables representing the three key abiotic aspects involved in plant development: bioclimatic, geo-physical, and edaphic aspects.

## 6.2. History of ELCmapas tool

The ELCmapas tool covered in this manual represents the development of the Ecogeographical Land Characterization Maps (ELC) concept published by Parra Quijano *et al.* (2012 A).

This type of maps has been put to diverse uses for the collection, conservation, and use of plant genetic resources as in the studies of Parra Quijano *et al.* (2011 A, 2011 B, and 2012 B) particularly for Spain, Thormann (2012) at the European level, Taylor *et al.* (2013) for the Czech Republic, Phillips *et al.* (2016) for Norway, Fitzgerald *et al.* (2019) for the Nordic countries, Contreras-Toledo *et al.* (2019) for Mexico, Tapia *et al.* (2019) in Ecuador, or Zair *et al.* (2020) for Middle Eastern countries.

The interest that this methodology prompted among various teams and research projects regarding the collection, conservation, and use of plant genetic resources contrasted with a specific observation made repeatedly by potential users. The methodology described in that publication was complex because it mixed geographic information systems (GIS) with multivariate analysis techniques. Also, the original development implied the use of a payment-based program to carry out statistical analysis. These issues were a major hindrance to the generation of ELC maps by researchers and technical experts.

## 6.3. Features of ELCmapas

ELCmapas tool is a new option that uses R to develop ELC maps and also avoids the complications described above. This free software can compute large amounts of statistical data and has an impressive array of graphic resources able to integrate GIS with multivariate analysis. The tool can produce ELC maps without switching between different programs or downloading and manipulating ecogeographic information. It is important to note that ELC tool products are maps and tables that can be visualized in programs such as DIVAGIS, Google Earth, or Microsoft Excel and, thus, these maps can be used as a component of other tools like Representa.

ELCmapas tool offers the user six different procedures to objectively determine the optimal number of groups to use in the clustering analysis. These procedures are:

- a)** elbow: A simple system that uses K - means as a clustering algorithm where the cut-off point is determined based on the decrease in the sum of the intra-group squares (Ketchen & Shook, 1996). The optimal number of groups is reached when the decrease in the intra-group sum of squares in a range of  $n$  and  $n+1$  groups is less than 50%. This is the fastest calculation method, also known as 'elbow', as it can process large amounts of data without long delays and is thus recommended for large countries.
- b)** medoides: Method of partition clustering around the medoids (pam). The method of silhouette interpretation and validation of the number of groups is used. This system (principally graphic, later adapted to R by the fpc package) allows determining how well data have been clustered (Kaufman and Rousseeuw, 1987; Rousseeuw, 1987). As this system consumes more computing resources, it takes considerably longer when applied to large data sets.
- c)** Calinski-Harabasz (1974) or calinski criterion to select the optimal number of clusters from applying the Kmeans method in 'cascade'. Like the 'medoides' method, an optimal number of cluster determination and

cluster assignment to each cell can demand a long time when the work frame is big (memory errors can be produced when using layers with more than 800,000 cells, for example, areas or countries higher than 800,000 km<sup>2</sup> at 1x1 km resolution).

- d) 'Simple structure index' (Donilcar *et al.*, 1999) or SSI criterion to select the optimal number of clusters from the application of the Kmeans method in 'cascade'. Like the 'medoides' method, an optimal number of cluster determination and cluster assignment to each cell can demand a long time especially when the work frame is big (memory errors can be produced when using layers with more than 800,000 cells, for example, areas or countries higher than 800,000 km<sup>2</sup> at 1x1 km resolution).
- e) 'Bayesian information criterion' or BIC. Under this criterion, the optimal number of clusters is established for parameterized Gaussian mixture models initialized by model-based hierarchical clustering. Since the detailed description of this criterion and the clustering method is unapproachable in this user manual, it is highly advisable to check the Fraley and Raftery (1998) study and their implementation of the method in the mclust package (see <https://cran.r-project.org/web/packages/mclust/index.html>). The application of this method is highly demanding of computing and memory capacity; therefore, it should not be used for work frames above 50,000 km<sup>2</sup> (1x1 km cell resolution) or 350,000 km<sup>2</sup> (5x5 km cell resolution).
- f) kmeansbic: It is a combination of the k means procedure with the determination of BIC. To do this, the function in R performs the clustering process used in the discriminant analysis of principal components (DAPC). The process consists of performing successive k means clustering procedures by using the principal components extracted from the original data as variables, where the number of clusters (k) increases in each cluster. For each k means, the Bayesian information criterion (BIC) is calculated as a statistical measure of goodness of fit, which allows determining the optimal number of clusters.

The methods used to determine the numbers of groups are not entirely objective, since the user decides the maximum number of groups allowed. As ecogeographic information at a resolution of 1 km or even 5 km for an entire subcontinent such as Latin America can be considerable, ELCmapas tool is best used at a country level. However, the distribution of the target species or the distribution of germplasm collections may exceed national borders. Therefore, ELC maps at regional (several contiguous countries), subcontinental, continental, and even global levels would be necessary. For this reason, the CAPFITOGEN3 version of ELCmapas tool would be capable of processing information and providing results at regional, subcontinental, continental, or global levels when used in a medium to high-performance computer or even a server.

## 6.4. Selecting ecogeographic variables

It is important to bear in mind that the selection of ecogeographic variables needs to be established before using ELCmapas tool. For this reason, SelecVar tool (chapter 5) was developed. Any change, addition, or deletion of a single variable of a single component (bioclimatic, geophysical, or edaphic) will significantly alter the final configuration of the map and its correlation with the adaptive scenarios of the species.

Originally ELC mapping techniques did not envisage a need for a higher level of discrimination between the variables, given that the objective was to create maps for general use. However, as their ability to discriminate correctly

between adaptive scenarios increased when focusing on a particular species or a group of closely related species (in genetic terms), we came up with the idea of selecting the ecogeographic variables of each component with the greatest influence on the abiotic adaptation of the species and that consequently determined their distribution.

The process used to select variables is critical to obtain more accurate maps in adaptive terms. The list of variables that can be potentially selected can be obtained from:

- a) Literature searches: It is easy to find references in technical and/or scientific publications (such as articles, books, or documents) about the environmental factors that influence, determine, or limit the distribution of a species. To make a map, the references on those factors should correspond to the variables in the form of GIS layers.
- b) Expert knowledge: Consultation with experts in the species or group of species often yields highly valuable information when selecting variables to know what ecogeographic variables are key for species' adaptation and distribution. Although the query introduces subjectivity into the process, this is not something to be afraid of. When creating ELC maps, resorting to expert knowledge during the preliminary stages can make the difference between a successfully validated map and a map with little meaning in terms of the target species' adaptation. The higher the number of experts consulted, the more decisive the contribution of expert knowledge to achieving an informed consensus. The work of Parra Quijano *et al.* (2012 C) is a good example of an ELC map created based on expert knowledge. In this study, the map was used to determine the ideal location of genetic reserves for several Beta species in Europe.

As an alternative to preparing this potential list of subjective variables, the user can determine which variables are of greatest importance when classifying the scenarios or which may be providing redundant information for each component (bioclimatic, geophysical, and edaphic). For this, an analysis of classification (such as Random Forest), an analysis of bivariate correlations or an analysis of collinearity can be performed. If high correlation values are found between two variables of the same component, one of them should be discarded. Furthermore, an analysis of the principal components (in case all variables are quantitative) can help to define the relationships between variables and determine the final selection. Some of these methods are available to users of the CAPFITOGEN technology in SelecVar tool (see Chapter 5). No more than five variables should be used per component since the configuration of the zones (adjacent cells with the same value) in the ensuing map may be difficult to read. Similarly, the use of latitude and longitude (parameters 6.5.1.7 and 6.5.1.8) results in maps with more geographically grouped categories and fewer crosslinks. The opposite effect is obtained by using variables such as 'orientation' from the geophysical component.

Once the final list of variables has been determined, these are selected in parameters bioclimv, geophysv, and edaphv.

## 6.5. Using ELCmapas tool

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and ELCmapas tool has been selected, the user should specify a series of parameters.



## **6.5.1 Initial parameters defined by the user**

### **6.5.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) to indicate the path to the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

### **6.5.1.2 Parameter: pais**

Explanation: Select the country for which you wish to build the ELC map. In the drop-down list, there are more than 160 countries and some sub-continental, global, and custom coverages (from rLayer tool).

### **6.5.1.3 Parameter: resol1**

Explanation: Select the degree of resolution you wish to use to generate the map. Note that 1x1 km offers greater resolution but requires greater computing capacity and takes far longer than 5x5 km, particularly in countries with a large land mass. High-resolution maps (i.e., 1x1 km) composed of more than 100,000 cells (countries with more than 100,000 km<sup>2</sup>) can generate processing problems for some methods of clustering and determination of the optimal number of groups such as 'cluster'.

### **6.5.1.4 Parameter: bioclimv**

Explanation: List (*local mode*) or select (*on server mode*) the bioclimatic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of bioclimatic variables'); copy the names and paste them separated by a semicolon (;). You will also find a blocked line with the 19 bioclim variables ready to use; simply remove the initial # symbol. In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'bioclimv'. To know the codes, names, and brief descriptions of the variables, consult the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

### **6.5.1.5 Parameter: geophysv**

Explanation: List (*local mode*) or select (*on server mode*) the geophysical variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of geophysical variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'geophysv'. To know the codes, names, and brief descriptions of the variables, consult the 'Variables names - Nombres de variables.xlsx' file downloadable

from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 6.5.1.6 Parameter: *latitud*

Explanation: If TRUE (*local mode*) or ✓ (*on server mode*) are set, the latitude (Y) will be included as a geophysical variable to be analysed.

#### 6.5.1.7 Parameter: *longitud*

Explanation: If TRUE (*local mode*) or ✓ (*on server mode*) are set, the longitude (X) will be included as a geophysical variable to be analysed.

#### 6.5.1.8 Parameter: *edaphv*

Explanation: List (*local mode*) or select (*on server mode*) the edaphic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of edaphic variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'edaphv'. To know the codes, names, and brief descriptions of the variables, consult the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 6.5.1.9 Parameter: *maxg*

Explanation: Please indicate the maximum number of clusters per component (bioclimatic, geophysical, and edaphic) that you wish to allow (the larger the number, the more categories on the map). We recommend values lower than or equal to five, otherwise, ELC maps of more than 125 categories can be generated.

#### 6.5.1.10 Parameter: *metodo*

Explanation: Select one of the methods offered to generate the clusters with an objective determination of the optimal number of clusters. The six methods described in this chapter are available by the terms: 'kmeansbic', 'medoides', 'elbow', 'calinski', 'ssi' and 'bic'.

#### 6.5.1.11 Parameter: *iterat*

Explanation: Applies only when you have selected 'calinski' or 'ssi' as methods in parameter 'metodo'. Indicate in this field the number of iterations used to generate K-means clusters to calculate 'calinski' or 'ssi' criteria.

### 6.5.1.12 Parameter: resultados (only for local mode)

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved (*local mode*). Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 6.6. Results of ELCmapas

After defining all the parameters and paths (for *local mode*) that ELCmapas requires, click on the 'Run' (*local mode* in RStudio) or 'Start' (*on server mode*) buttons to start the analysis process of the tool.

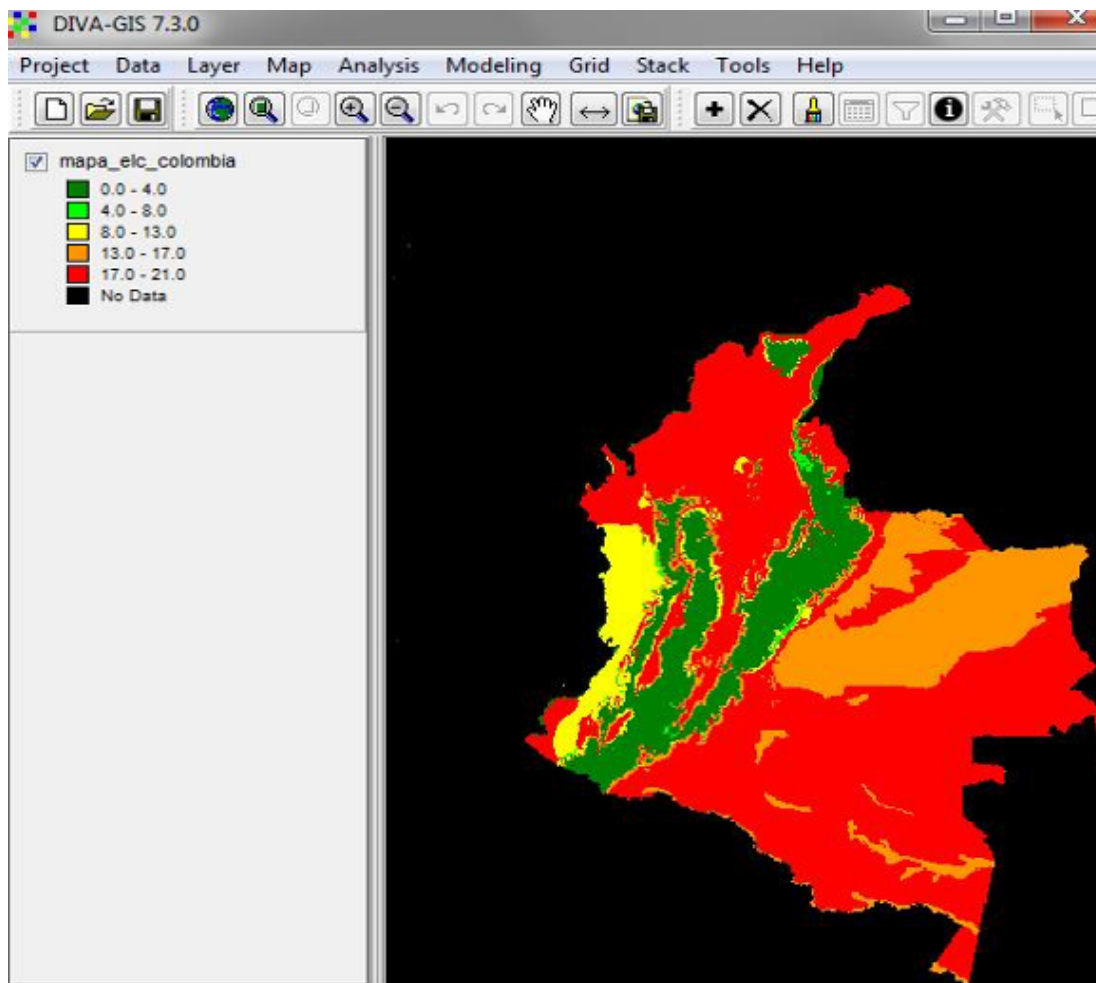
Note that the waiting time will vary according to the type of analysis run. The results obtained by ELCmapas will be automatically saved in the path and folder previously typed in "resultados" (in *local mode*) or in the User's Files and Results area in *on server mode*. There, the user will find five maps and three tables.

### 6.6.1 The maps

The final ELC map of the country or region determined in the parameter of section 6.5.1.2 is contained in the mapa\_elc\_pais.tif (geotif) file. It can also be viewed and edited in DIVA-GIS as a grd version (when opening the mapa\_elc\_DIVA\_pais.grd file), in other graphic formats, or for Google earth (png or kml respectively). If the resulting ELC map is going to be used by other tools such as Representa, ColNucleo, FIGS\_R or Tzones, you must copy the mapa\_elc\_pais.grd file (which should not be opened or edited) and paste it in the corresponding folder (*local mode*), or you must upload it (*on server mode*) through the button that enables the selection and subsequent uploading of files.

The maps that represent the resulting categories from the bioclimatic, geophysical, and edaphic components (mapa\_bioclimatico\_pais.tif, mapa\_geofisico\_pais.tif, and mapa\_edafico\_pais.tif) can be opened in different GIS software. However, you must remember that these maps are in the WGS84 coordinate system (latlong). All maps have an alternative version in .grd, which is the native format for the DIVA-GIS software. Initially, DIVA-GIS opens the maps as shown in Fig. 29.

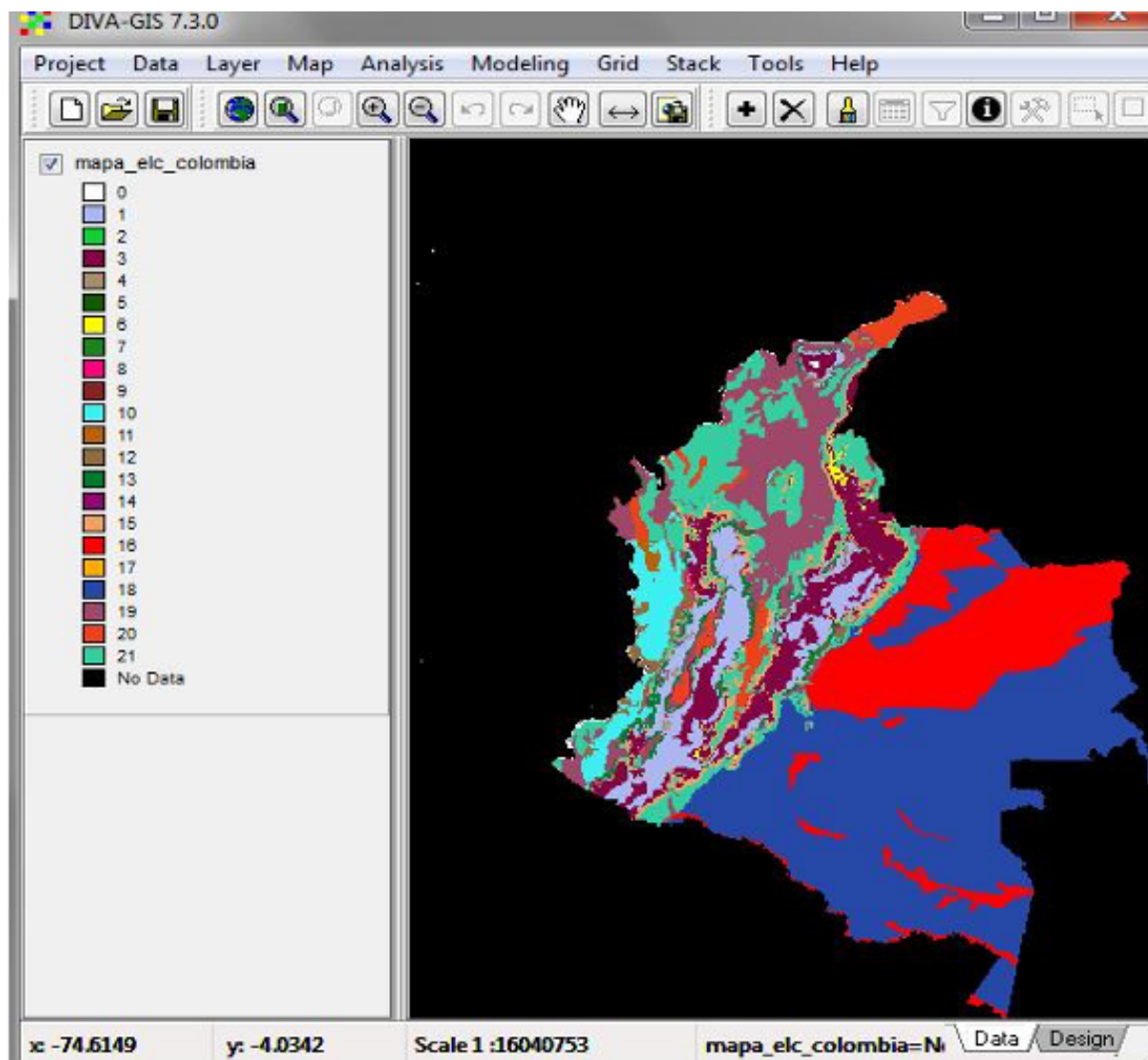
However, the display may be altered by double-clicking on the grey panel on the left that represents this layer in the DIVA-GIS interface. By adding as many rows as there are categories in the map, and then applying a swatch of random colours, you can obtain a map such as that shown in Fig. 30. It helps to use widely contrasting colours so that the categories (ecogeographic scenarios) present in the territory may be easily identified.



**Figure 29.** Example of a map of Colombia generated by ELCmaps tool as pictured in DIVA-GIS before any editing. The 20 categories are shown grouped into 5 ranges.

**NOTE:** Always remember that the '0' (zero) category is not one of the ecogeographic categories in the final map; rather, it is used to refer to those areas for which there is information for one or two components but not all three. For example, for obvious reasons, there is no soil information for urban areas or bodies of water, but there may be information on bioclimatic and even geophysical features for these areas. Those areas will be coded as '0'.

In addition to the .tif and .grd maps, a map compatible with Google Earth ('mapa\_elc\_pais.kml') is produced. If you have Google Earth installed on your computer, this map may be opened as a layer over the Google Earth images when you double-click on the file. This map may not be manipulated (to change the colours) and it does not have optimal graphic quality.



**Figure 30.** Example of a map of Colombia generated by the ELCmapas tool, showing a colour for each category. The properties of the map as opened by DIVA-GIS have been altered to show each category in a different colour.

### **6.6.2 The Tables**

The tables produced by ELCmapas and the rest of the CAPFITOGEN3 applications are offered in .txt and .xls extensions. While this last extension corresponds to Microsoft Excel®, it is the same tab-delimited text file. So, why duplicate the same information in the same format offering it with two different extensions? The explanation is that the .xls extension indicates the operating system that the table must be opened in Excel®. When trying to open this text file, Excel® recognizes that it does not correspond to its format and emits a warning message (see Fig. 25), which can be ignored. The .txt format can be opened with the option 'Open with' and the next step would be indicating some compatible software, including Excel® itself.

**6.6.2.1 'Tabla\_ELC\_celdas\_pais.txt'**. This table shows the values of the selected variables and the values of the ELC categories ('ELC\_CAT'), which are the bioclimatic, geophysical, and edaphic categories for each cell centroid (row) making up the territory of the country under study. It also includes latitude and longitude values for each centroid.

**6.6.2.2 'numero\_categorias\_pais.txt'**. This table contains a simple count of the ecogeographic categories that have been generated and represented in the resulting ELC map (column 'N\_ELC\_CAT') and the number of categories generated by each component.

**6.6.2.3 'Estadist\_ELC\_pais.txt', 'Estadist\_BIOCLIM\_pais.txt', 'Estadist\_EDAPH\_pais.txt', 'Estadist\_GEOPHYS\_pais.txt' and 'Combi\_ELC\_ecuador.xls'**. These tables include the descriptive statistics (mean, minimum value, maximum value, and standard deviation) for each of the original variables involved in the creation of the ELC map and for the maps of each of the components (bioclimatic, geophysical, and edaphic) represented on the ELC map. These tables are similar to the S2 supplementary table shown to describe the categories of the ELC map in the paper by Parra Quijano *et al.* (2012a). This table is required together with the `mapa_elc_pais.grd` file and its accompanying `mapa_elc_pais.gri` file when working with Representa and ColNucleo tools (FIGS\_R optionally) and must be copied together with the map files in the folder indicated by each tool (*local mode*) or uploaded in parameter 'statelc' (*on server mode*).

**6.6.2.4 'Bioclim\_BIC\_results.txt', 'edaph\_BIC\_results.txt', and 'edaph\_BIC\_results.txt'**. These tables result from the selection of the 'bic' method in parameter 6.5.1.10. They contain the most important parameters of the clustering analyses for each component, such as the model and the number of optimal clusters used to generate the ELC map. A graph in jpg format is also generated to show the BIC values vs. the number of clusters for each component (bioclimatic, edaphic, or geophysical).

## 6.7. References

- Calinski, T., Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics*. 3(1): 1-27.
- Contreras-Toledo, A. R., Cortés-Cruz, M., Costich, D. E., Rico-Arce, M., Magos Brehm, J., Maxted, N. 2019. Diversity and conservation priorities of crop wild relatives in Mexico. *Plant Genetic Resources Characterisation and Utilisation*, 17: 140-150.
- Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., Hahn, N., Palminteri, S., Hansen, M. 2017. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6): 534-545.
- Dolnicar, S., Grabler, K., Mazanec, J. A. 1999. A tale of three cities: perceptual charting for analyzing destination images. Pp. 39-62 *In*: Woodside, A. *et al.* (eds.), *Consumer psychology of tourism, hospitality and leisure*. CAB International, New York.
- Fitzgerald, H., Palmé, A., Asdal, Å., Endresen, D., Kiviharju, E., Lund, B., Rasmussen, M., Thorbjornsson, H., Weibull,

- J. 2019. A regional approach to Nordic crop wild relative in situ conservation planning. *Plant genetic resources*, 17(2): 196-207.
- Fraley C., Raftery, A.E. 2007. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification* 24:155-181.
- Kaufman, L., Rousseeuw, P.J. 1987, Clustering by means of Medoids, in *Statistical Data Analysis Based on the L1-Norm and Related Methods*. Y. Dodge (ed.), North-Holland, 405–416.
- Ketchen, D. J., Shook, C. L. 1996. The application of cluster analysis in Strategic Management Research: An analysis and critique. *Strategic Management Journal* 17(6): 441-458.
- Parra-Quijano, M. Iriondo, J.M., De la Cruz, M., Torres, M.E. 2011 A. Strategies for the development of core collections based on ecogeographical data. *Crop Science* 51:656-666
- Parra-Quijano, M. Iriondo, J.M., Torres, M.E., De la Rosa, L. 2011b. Evaluation and validation of ecogeographical core collections using phenotypic data. *Crop Science* 51: 694-703.
- Parra-Quijano, M.; Draper, D.; Torres, E., Iriondo, J.M. 2008. Ecogeographical representativeness in crop wild relative *ex situ* collections. p. 249-273. *In: Maxted, N.; Ford-Lloyd, B.V.; Kell, S.P.; Iriondo, J.M.; Dulloo, M.E., Turok, J. (eds.) Crop wild relative conservation and use*. CAB International, Wallingford.
- Parra-Quijano, M. Iriondo, J.M., Torres, M.E. 2012a. Ecogeographical land characterization maps as a tool for assessing plant adaptation and their implications in agrobiodiversity studies. *Genetic Resources and Crop Evolution* 59(2): 205-217. doi: 10.1007/s10722-011-9676-7
- Parra-Quijano, M. Iriondo, J.M., Torres, M.E. 2012b. Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps. *Biodiversity and Conservation* 21: 79-96. doi: 10.1007/s10531-011-0167-0
- Parra-Quijano, M. Iriondo, J.M., Frese, L., Torres, M.E. 2012c. Spatial and ecogeographic approaches for selecting genetic reserves in Europe. *In: Maxted, N., Dulloo, M.E., Ford-Lloyd, B.V., Frese, L., Iriondo, J., Pinheiro de Carvalho, M.A.A. (eds.) Agrobiodiversity Conservation: securing the diversity of crop wild relatives and landraces*. CABI, Wallingford, UK.
- Phillips, J., Asdal, Å., Magos Brehm, J., Rasmussen, M., Maxted, N. 2016. *In situ* and *ex situ* diversity analysis of priority crop wild relatives in Norway. *Diversity and Distributions*, 22(11): 1112-1126.
- Rousseeuw, P.J. 1987. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53-65. doi: 10.1016/0377-0427(87)90125-7.

Tapia, C., Paredes, N., Lima, L. (2019). Representatividad de la diversidad del género musa en el Ecuador. Revista Científica Ecuatoriana, 6(1).

Taylor, N. G., Kell, S. P., Holubec, V., Parra-Quijano, M., Chobot, K., Maxted, N. (2017). A systematic conservation strategy for crop wild relatives in the Czech Republic. Diversity and Distributions, 23(4): 448-462.

Thormann, I. 2012. Applying FIGS to crop wild relatives and landraces in Europe. Crop Wild Relative 8 14:16.

Tohme, J., Jones, P., Beebe, S., Iwanaga, M. 1995. The combined use of agroecological and characterisation data to establish the CIAT *Phaseolus vulgaris* core collection. p. 95-107. In Hodgkin, T., Brown, A.H.D., van Hintum, Th.J.L., Morales, E.A.V. (eds.) Core collections of plant genetic resources. IPGRI, Rome.

Zair, W., Maxted, N., Brehm, J. M., Amri, A. 2020. *Ex situ* and *in situ* conservation gap analysis of crop wild relative diversity in the Fertile Crescent of the Middle East. Genetic Resources and Crop Evolution, 1-17.







**Birmingham Workshop (UK), February 2014.**



# 7 | ECOGEO Tool

## 7.1. Ecogeographic Characterization of Germplasm

Ecogeographic characterization is understood as the analysis of all environmental information from the growth site of an individual plant or plant population, directly related to the process of adaptation to the biotic or abiotic environment. CAPFITOGEN tools only analyse the abiotic component, which is classified according to three principal features that are often considered in studies of crop adaptation (Ceballos-Silva & Lopez-Blanco, 2003) and agricultural zoning (Williams *et al.*, 2008):

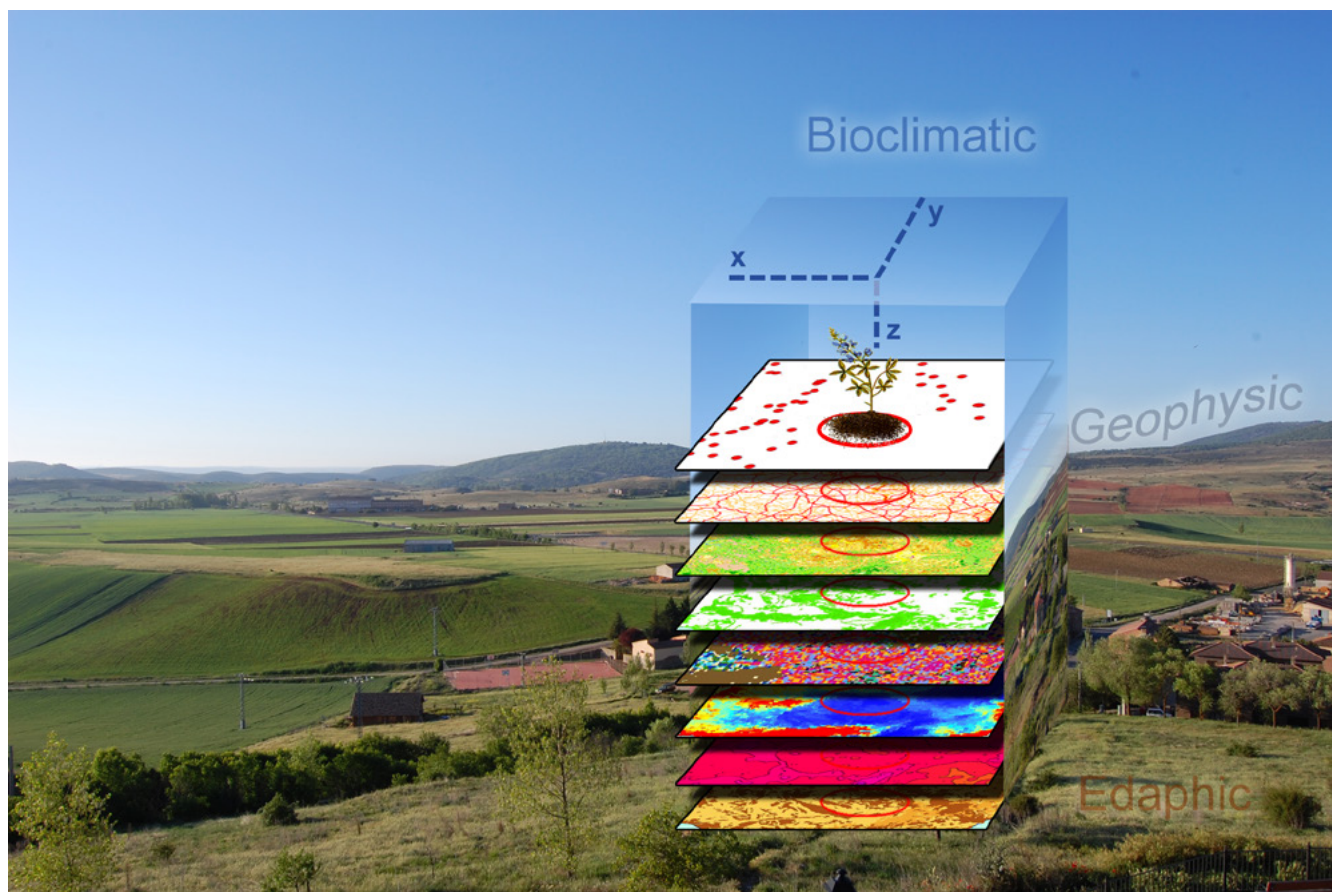
- a) **Bioclimatic:** This refers to factors related to temperature and rainfall. It also includes the relationships between temperature and rainfall that are managed using indexes.
- b) **Geophysical:** This brings together topographical and relevant relief factors, especially those related to solar radiation.
- c) **Edaphic:** This concerns factors related to the physical and/or chemical conditions of the soil on which plants depend.

Therefore, the ecogeographic characterization of a set of accessions involves assigning the bioclimatic, geophysical, and edaphic information from the collection site to each accession.

Ecogeographic information from a collection site reveals many adaptive traits of the germplasm. If considered in conjunction with other types of characterizations, such as phenotypic or genotypic, it can be very useful in explaining the genetic patterns observed. In cases where economic resources are too scarce for other kinds of studies, ecogeographic characterization is a valid, simple, and inexpensive alternative to using germplasm for breeders seeking parent plants with certain adaptive traits in the collections.

The most important input required for an ecogeographic characterization is the collection site's coordinates or its description (from which the coordinates may be extracted), usually recorded in the passport descriptors at the time of collection. Using these coordinates, data may be assigned to each accession describing the most important environmental features of the collection site. The quality of these coordinates is thus a crucial aspect for the proper allocation of ecogeographic information, which is why GEOQUAL tool should be used before performing a characterization of this type. In addition to the coordinates as raw material, ecogeographic characterization requires environmental information about the entire work area as well as a GIS project management software to extract the information corresponding to each collection site.

The product of an ecogeographic characterization is similar to other types of characterization: it is a data matrix where the rows usually correspond to the accessions and the columns to the descriptors. From this initial matrix, it is possible to perform multivariate analyses, such as those frequently performed with other types of characterization, to determine the environmental similarity between different collection sites. One such factorial analysis, (for example, the Principal Component Analysis (PCA)), would also highlight the relationship between the different variables originally entered and create synthetic noncorrelated variables describing the ecogeographic affinities between the inputs with a reduced number of components.



**Figure 31.** Process used to extract ecogeographic information for a collection site using GIS.

Please note that ecogeographic characterization yields information about the collection sites, rather than the nature of the germplasm itself. Therefore, multivariate analyses that operate on matrices of distance or dissimilarity here reflect the environmental affinity and, indirectly, the adaptive affinity between different collection sites. Accessions for the same species with different genotypic or phenotypic patterns may occur in very similar or even indistinguishable environmental scenarios.

## 7.2. Features of ECOGEO

ECOGEO tool provides the user with ecogeographic information extracted from over 160 variables for a list of accessions to be entered into the analysis using the format for passport data FAO/Bioversity 2012, with minor modifications. However, from CAPFITOGEN3, ECOGEO can perform multivariate analyses with phenotypic or genotypic characterization data in the same way as DIVmapas tool. These analyses are an addition to the complete ecogeographic characterization process the tool already performed. This makes ECOGEO a much more complete tool for technicians who conserve and characterize plant genetic resources for food and agriculture. If you have any questions about the characteristics of the phenotypic or genotypic characterization tables, you can find more details about them in the chapter on DIVmapas tool.

CAPFITOGEN offers the user all the necessary information for an ecogeographic characterization with ECOGEO, avoiding the need to download, adapt and standardize layers of information from different internet sites. Thus, CAPFITOGEN's layers of ecogeographic variables are ready to use and work with the R program settings of ECOGEO tool, among others (see Chapter 2).

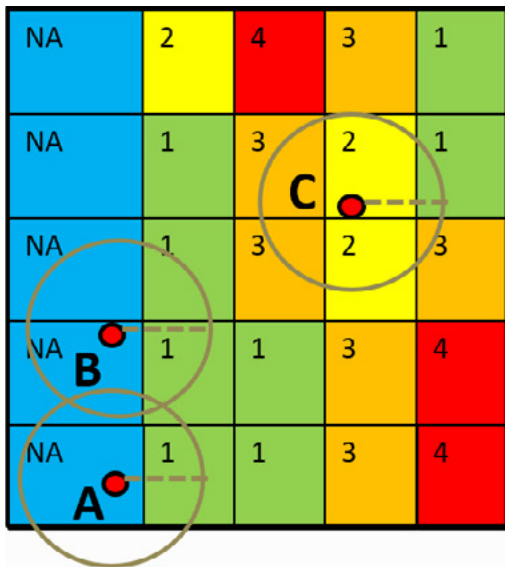
The work area is the second aspect defined by the user and often corresponds to national territorial boundaries as defined in the global database on administrative areas (<http://www.gadm.org>). Variables or ecogeographic layers are cut according to their limits, so that if a particular country is chosen but the passport data includes coordinates corresponding to sites outside the country, the accessions for these coordinates will not be assigned any information. Options may be available to draw up multi-country analyses for a given region or even a continent. If regions or continents are available, (appearing in the listings under parameter 'country'), the user can work with these areas of greater coverage, considering that the level of resolution of this information will probably be in a shorter range (cell sizes over 10x10 km).

There is an important aspect that ECOGEO tool introduces regarding the way in which ecogeographic information is extracted from a collection site. Extractions are usually performed at the point indicated by the coordinates. However, there are two situations when the 'punctual' extraction does not reflect the true nature of the abiotic conditions of the collection site:

- a) When there is little information available about the coordinates or these are poor quality, according to GEOQUAL or other methodologies. For instance, in cases of species with a coastal distribution, where, despite relatively accurate georeferencing, a specific extraction may yield several 'NA' values (no information available) because the ecogeographic information raster maps/layers do not perfectly fit the contours of the shoreline.
- b) When for various reasons the germplasm collection site does not correspond precisely to the site where the plant grows but is found within a relatively well-known perimeter (for example when germplasm is collected in local markets).

In these cases, the user may use a 'circular' extraction and provide the radius around the point indicated by the coordinates for which the information is to be extracted. Thus, ECOGEO extracts ecogeographic data from the full range of cells within the radius, calculates its average value, and assigns this value to the accession, repeating this process for all the ecogeographic variables used to characterize the germplasm (Fig. 32). Additionally, ECOGEO automatically discards 'NA' values when calculating and subsequently assigning values. To program ECOGEO to perform a 'circular' extraction, the user must first activate parameter 'buffy' and then enter the value in meters of the radius of the circular extraction area in parameter 'tamp'. This procedure of circular extraction is also available for other tools.

Once the user prepares the passport and phenotypic and/or genotypic characterization tables (if applicable) according to the pre-established format, they must indicate the tool where these tables are located (*local mode*) or provide them to the tool (*on server mode*). Then, the user must program the tool with the location and indicate the work area, resolution, and extraction method required. Finally, the only task remaining is defining the variables/layers of interest for each aspect (bioclimatic, geophysical, and edaphic) to perform the ecogeographic characterization of germplasm collection sites.



ACCENUMB	Captured Values*	Circular extraction (mean)	Punctual extraction
A	NA,1,1	1	NA
B	NA,1,1	1	NA
C	3,2,1,3,2,3	2.333	2

\*from centroids falling within circular areas

**Figure 32.** Differences between the values assigned from a specific extraction point and a circular extraction. Cells in blue and NA values represent bodies of water, while the red points indicate the three collection sites (identified using ACCENUMB codes) located from their coordinates.

With the definition of these parameters, in a single step, ECOGEO tool can seek out variables/layers of ecogeographic information of interest, group them and extract information for each coordinate from the group of layers. The information extracted is used to generate a table that will be saved wherever defined by the user in parameter 'resultados'.

Finally, if the user is interested in performing a cluster analysis or a Principal Component Analysis (PCA), the tool can be programmed to run these analyses. The type of grouping and the number of principal components to be retained may also be indicated at this point. ECOGEO tool will produce graphs (dendrograms or biplots) and tables (values and main vectors and scores for the retained components) which will be saved in the folder indicated in parameter 'resultados'. If phenotypic or genotypic analyses have been requested in addition to the ecogeographic ones, ECOGEO will perform the multivariate analyses and the results will also be saved in 'resultados'.

### 7.3. Using ECOGEO tool

Once the user has installed CAPFITOGEN3 *local mode* tools or accessed CAPFITOGEN3 *on server mode* and GEOQUAL tool has been selected, a set of parameters to ensure the R program runs correctly must be defined.

### **7.3.1 Initial parameters defined by the user**

#### **7.3.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### **7.3.1.2 Parameter: pais**

Explanation: Select the country/region where all or most of the data accessions you wish to analyse were collected. If accessions have been collected from more than one country, you may select a region, subcontinent, or continent. You can also use rLayer tool to produce your own work frames and select them here.

#### **7.3.1.3 Parameter: pasaporte**

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

#### **7.3.1.4 Parameter: geoqual**

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### **7.3.1.5 Parameter: totalqual**

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).



### 7.3.1.6 Parameter: buffy

Explanation: Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish ecogeographic information to be extracted from a circular area around the collection site. Leaving this option unchecked means that information is extracted only from the point indicated by the collection site coordinates. This option is very useful, for example, when most of the collection sites correspond to markets and not directly to the growers' farms, or when the actual location of the collection sites is uncertain even though GEOQUAL can grant high georeferencing quality.

### 7.3.1.7 Parameter: tamp

Explanation: Applies only if 'buffy' has been set as TRUE (✓ in *on server mode*). Specify the radius (in km) of a circular area around the point indicated by the collection site coordinates from which the ecogeographic information is to be extracted. The values extracted from the circular area will be averaged to obtain a single value. The information will be extracted from those cells whose centroid is within the circular area.

### 7.3.1.8 Parameter: resol1

Explanation: Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution or precision regarding some coordinates (points X and Y) but requires greater computing capacity than 5x5 km. Resolutions of 10x10 and 20x20 km may only be used for large countries, subcontinents, or continents.

### 7.3.1.9 Parameter: ecogeo

Explanation: Indicate here (with the TRUE option in *local mode* or ✓ in *on server mode*) if you are interested in carrying out an ecogeographic characterization of the germplasm collection/observation sites.

### 7.3.1.10 Parameter: bioclimsn

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to characterize by bioclimatic variables (annual or monthly average, maximum or minimum temperatures, monthly or annual rainfall, vapor pressure, etc.).

### 7.3.1.11 Parameter: bioclimv

Explanation: Applies only if 'bioclimsn' has been set as TRUE (✓ in *on server mode*). List (*local mode*) or select (*on server mode*) the bioclimatic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of bioclimatic variables'); copy the names and paste them separated by a semicolon (;). You will also find a blocked line with the 19 bioclim variables ready to use; simply remove the initial # symbol. In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will

appear in the box in front of parameter 'bioclimv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 local mode.

#### 7.3.1.12 Parameter: edaphsn

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to characterize the information by soil variables (texture, depth, pH, etc.).

#### 7.3.1.13 Parameter: edaphv

Explanation: Applies only if 'edaphsn' has been set as TRUE (✓ in *on server mode*). List (*local mode*) or select (*on server mode*) the edaphic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of edaphic variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'edaphv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 7.3.1.14 Parameter: geophysn

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to characterize the information by geophysical variables (related to terrain and solar radiation).

#### 7.3.1.15 Parameter: geophysv

Explanation: Applies only if 'geophysn' has been set as TRUE (✓ in *on server mode*). List (*local mode*) or select (*on server mode*) the geophysical variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of geophysical variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'geophysv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 7.3.1.16 Parameter: latitud

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) to include the latitude (Y) as a geophysical variable to be analysed.

### 7.3.1.17 Parameter: longitud

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) to include the longitude (X) as a geophysical variable to be analysed.

### 7.3.1.18 Parameter: phenotip

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a phenotypic characterization analysis. For this, you must have access to data for phenotypic characterization or evaluation (morphology, phenology, productivity, resistance, etc.) in the specified format. Please remember to include the name of the extension. For example, if the table is called 'phenotypes', in this space you must write 'phenotypes.txt'. Remember that this table must be located in the 'Pasaporte' folder of CAPFITOGEN3 *local mode* tools data structure. For *on server mode*, the table will be uploaded at the moment of configuring the tool or before this process so that it can be available in the User's Files and Results area.

### 7.3.1.19 Parameter: phenot

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Indicate the name of the text file containing the data from the phenotypic characterization in the format indicated in *local mode*. Please remember to include the name of the extension. For example, if the table is called 'phenotypes', in this space you must write 'phenotypes.txt'. For *on server mode*, you must click on the button that allows you to search, select and upload the corresponding file.

### 7.3.1.20 Parameter: phenotv

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Indicate the name of the text file containing the table describing the nature of each phenotypic variable in the format indicated for *local mode*. Please remember to include the name of the extension. For example, if the table is called 'phenotypevariables', in this space you must write 'phenotypevariables.txt'. This table must describe all the variables included in the table with the characterization data (see the previous parameter). For *on server mode*, you must click on the button that allows you to search, select, and upload the corresponding file.

### 7.3.1.21 Parameter: genotip

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a genotypic characterization analysis. For this, you must have access to data for phenotypic characterization or evaluation (such as the presence or absence of markers as zero and one) in the specified format. Remember that this table must be located in the 'Pasaporte' folder (*local mode*) of the CAPFITOGEN3 tools data structure. For *on server mode*, the table will be uploaded at the moment of configuring the tool or before this process so that it can be available in the User's Files and Results area.

**7.3.1.22 Parameter: *genot***

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Indicate the name of the text file containing the genotypic characterization data in the format indicated for *local mode*. Please remember to include the name of the extension. For example, if the table is called 'genotypes', in this space you must write 'genotypes.txt'. For *on server mode*, you must click on the button that allows you to search, select, and upload the corresponding file.

**7.3.1.23 Parameter: *neigd***

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Select this option if you wish to obtain a map of Nei's average index of genetic diversity (1987), a map of the average proportion of polymorphic markers, and a map of the number of accessions analysed by cell.

**7.3.1.24 Parameter: *csimilar***

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Indicate the similarity coefficient that you want to use to generate the map of average genotypic distance. 1 = Jaccard Index (1901), 2 = Simple Matching Coefficient (SMC) Sokal & Michener (1958), 3 = Sokal & Sneath (1963) (S5 coefficient of Gower & Legendre), 4 = Rogers & Tanimoto (1960), 5 = Dice (1945), 6 = Hamann coefficient, 7 = Ochiai (1957), 8 = Sokal & Sneath (1963) (S13 coefficient of Gower & Legendre), 9 = Pearson Phi coefficient, 10 = S2 coefficient of Gower & Legendre. Distance (d) is obtained as  $d = \sqrt{1-s}$  where s is the similarity coefficient.

**7.3.1.25 Parameter: *ecogeoclus***

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a cluster analysis for accessions with an ecogeographic characterization.

**7.3.1.26 Parameter: *ecogeoclustype***

Explanation: Applies only if 'ecogeoclus' has been set as TRUE (✓ in *on server mode*). Choose the type of hierarchical cluster to be used for ecogeographic clusters: 'single' = nearest neighbour, 'complete' = more compact neighbourhood, 'ward' = method of minimum variance of Ward, 'mcquitty' = McQuitty's method, 'average' = average similarity (UPGMA), 'median' = similarity of the median, 'centroid' = geometric centroid, 'flexible' = flexible Beta.

**7.3.1.27 Parameter: *ecogeopca***

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to perform an analysis of principal components for accessions with an ecogeographic characterization.

**7.3.1.28 Parameter: ecogeopcaxe**

Explanation: Applies only if 'ecogeopca' has been set as TRUE (✓ in *on server mode*). Indicate here the number of components to retain within the PCA analysis. This number should always be lower than the number of ecogeographic variables. This means that only the 'retained' principal components will be shown in the results tables.

**7.3.1.29 Parameter: phenoclus**

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform a cluster analysis of all accessions that include phenotypic information.

**7.3.1.30 Parameter: phenoclustype**

Explanation: Applies only if 'phenoclus' has been set as TRUE (✓ in *on server mode*). Choose the type of hierarchical cluster to be used for phenotypic clusters: 'single' = nearest neighbour, 'complete' = more compact neighbourhood, 'ward' = method of minimum variance of Ward, 'mcquitty' = McQuitty's method, 'average' = average similarity (UP-GMA), 'median' = similarity of the median, 'centroid' = geometric centroid, 'flexible' = flexible Beta.

**7.3.1.31 Parameter: phenopca**

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform a principal component/coordinate analysis of all accessions that include phenotypic information.

**7.3.1.32 Parameter: phenopcaxe**

Explanation: Applies only if 'phenopca' has been set as TRUE (✓ in *on server mode*). Indicate here the number of components/coordinates to retain within the PCA/PCoA analysis. This number should always be lower than the number of phenotypic variables. This means that only the 'retained' principal components will be shown in the results tables.

**7.3.1.33 Parameter: phenovarq**

Explanation: Applies only if 'phenopca' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if all the phenotypic variables/descriptors correspond to quantitative variables.

**7.3.1.34 Parameter: genoclus**

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform a cluster analysis of all accessions that include genotypic information.

**7.3.1.35 Parameter: genoclustype**

Explanation: Applies only if 'genoclus' has been set as TRUE (✓ in *on server mode*). Choose the type of hierarchical cluster to be used for genotypic clusters: 'single' = nearest neighbour, 'complete' = more compact neighbourhood, 'ward' = method of minimum variance of Ward, 'mcquitty' = McQuitty's method, 'average' = average similarity (UPG-MA), 'median' = similarity of the median, 'centroid' = geometric centroid, 'flexible' = flexible Beta.

**7.3.1.36 Parameter: genopco**

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform an analysis of the principal coordinates of all the accessions that include genotypic information.

**7.3.1.37 Parameter: genopcoaxe**

Explanation: Applies only if 'genopca' has been set as TRUE (✓ in *on server mode*). Indicate here the number of coordinates to be retained within the PCoA analysis. This number should always be lower than the number of genotypic variables or markers. This means that only the 'retained' principal coordinates will be shown in the results tables.

**7.3.1.38 Parameter: mantelt**

Explanation: Indicate (with TRUE in *local mode* or ✓ in *on server mode*) if you wish to analyse the correlation matrix (Mantel, 1967) between the possible combinations of factors (ecogeographic vs. phenotypic vs. genotypic). All possible comparisons will be made according to whether phenotypic or genotypic data were entered or if an ecogeographic characterization matrix was created based on collection sites. A matrix of geographic distances will be generated for paired matrix comparisons.

**7.3.1.39 Parameter: mantelmeth**

Explanation: Applies only if 'mantelt' has been set as TRUE (✓ in *on server mode*). Select the type of correlation to use in the Mantel test.

**7.3.1.40 Parameter: mantelper**

Explanation: Applies only if 'mantelt' has been set as TRUE (✓ in *on server mode*). Enter the number of permutations you want to perform the Mantel test.

**7.3.1.41 Parameter: resultados (only for local mode)**

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / instead of \ when indicating the path of the folder. For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 7.4. Results of ECOGEO

After defining all the parameters and paths (for *local mode*) that ECOGEO requires, click on the 'Run' (*local mode* in RStudio) or 'Start' (*on server mode*) buttons to start the analysis process of the tool.

After some time that may vary due to the type of analysis requested, ECOGEO will save the results in the path and folder specified in 'resultados' (in *local mode*), or in the User's Files and Results area in *on server mode*.

Regardless of this, ECOGEO can generate up to three different folders with results and a file with the parameters used to generate these results (Parameters.Parameters.ECOGEO.txt). The contents of the folders are detailed below.

### 7.4.1 ClassicMultivariateResults pais folder

#### 7.4.1.1 Figures

These are files called dendrograma\_ecogeo/fenotípico/genotípico.wmf and pca\_ecogeo.wmf or pco\_fenotípico/genotípico.wmf; they are vector figures in Windows Metafile format. The figures (a dendrogram and a biplot) are only generated if the tool has been instructed to perform cluster analysis or an analysis of principal components (parameters ecogeoclus, ecogeopca, phenoclus, phenopca, genoclus, and genopco). They may be opened and even modified in Microsoft PowerPoint or image editing programs.

#### 7.4.1.2 Tables

Tables are generated as a result of the analysis of principal components. They correspond to the following files: ecogeographic\_eigenvalues.txt/xls (table of eigenvalues), ecogeographic\_eigenvectors.txt/xls (table of eigenvectors), ecogeographic\_pcascotes.txt/xls (table containing each accession's score for the principal components retained) and/or phenotypic/genotypic\_pcascotes.txt/xls (table containing each accession's score for the principal coordinates retained). These tables are only generated if the tool has been required to perform the analyses of principal components or coordinates.

### 7.4.2 EcogeographicCharacterization pais folder

**7.4.2.1 Ecogeographic characterization table of the accessions:** This is the file called TablaVarEcogeograficapais.txt/xls. It corresponds to the initial characterization matrix and contains as many rows as accessions analysed and as many columns as ecogeographic descriptors.

### **7.4.3 MantelCorrelationResults\_pais folder**

All tables with the distance matrices calculated for all accessions simultaneously ('Matriz\_distancia\_') and those containing the results of Mantel's matrix correlation tests (1967) will be saved in this folder. The name of each table indicates the kind of comparison process made. Dice's distance matrix is used to measure correlations where genotypic data are involved. For example, the file 'Mantel\_genotypic\_Vs\_phenotypic.txt' contains the results of the correlation matrix between genotypic distances (Dice) and phenotypic distances (Gower). It is important to note that ECOGEO also calculates the matrix of geographical distances (calculated in decimal degrees) to enable matrices to be compared in terms of the geographical distance component.

## **7.5. References**

Ceballos-Silva, A., López-Blanco, J. 2003. Evaluating biophysical variables to identify suitable areas for oat in Central Mexico: a multi-criteria and GIS approach. *Agriculture, Ecosystems and Environment* 95:371–377.

Williams, C.L., Hargrove, W.W., Liebman, M., James, D.E. 2008. Agro-ecoregionalization of Iowa using multivariate geographical clustering. *Agriculture, Ecosystems and Environment* 123:161–174







**National Workshop, Quito (Ecuador), August 2017.**



# 8 | Representa Tool

## 8.1. Concept of representativeness in germplasm collections

Certain sensitive issues may jeopardize the successful *ex situ* conservation of plant genetic resources. These may arise at two specific moments: at the time of collection or during conservation *per se*. The risk of losing accessions during the conservation period may be reduced by applying appropriate techniques to manage germplasm. Nonetheless, the germplasm selected for conservation must be the most faithful reflection possible of the genetic diversity of plant populations occurring in the field. In the best-case scenarios, this reflection should remain intact without the need for new collections. This situation highlights the importance of collecting germplasm in a manner that ensures the capture of the broadest genetic diversity possible. The **representativeness** of a germplasm collection measures the ability of the conserved sample to represent the full range of genetic diversity occurring in nature.

The representativeness of a species in a germplasm collection can be determined at the intra- and inter-population levels. In the case of a cultivated species, the equivalent would be the intra- and inter-varietal levels. These two concepts are inseparable when taking the representativeness of a collection as a whole. Despite this, and due to practical issues related to the way in which germplasm conservation is carried out, both concepts have hitherto been worked independently of each other.

The intra-population representativeness has been exhaustively studied, as in the multiple papers by Crossa *et al.* (1994, 1997, 2011), which has resulted in the design of specific collection strategies according to the reproductive biology of the species, the spatial distribution of the individuals, and the size of the population. The idea is to calculate on a case-by-case basis the minimum number of individuals to be collected to ensure the capture of most of the alleles present in the population. In contrast, there has been less work on how to represent a species in a collection in inter-population terms. However, since the development of the concept of core collections, the inter-population representation of a species in a collection has gained importance, given that these subcollections only operate at this level (Brown, 1989; Yonezawa *et al.*, 1995).

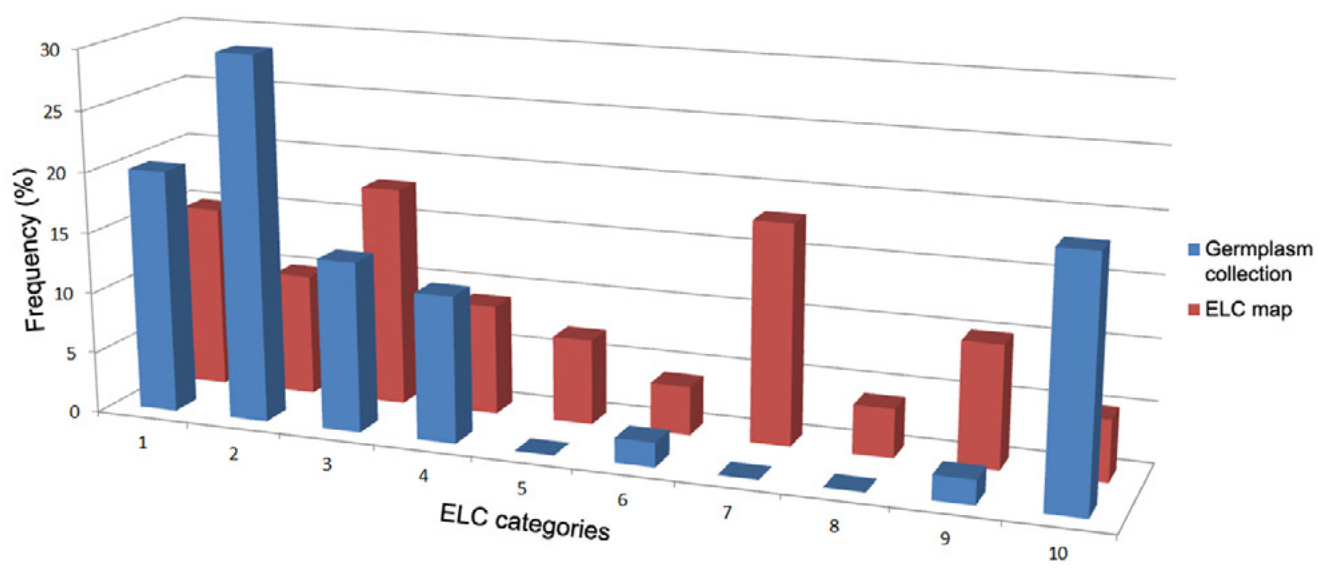
Once the concept of representativeness of a germplasm collection had taken root in the community of scientists and curators working in the field of plant genetic resources, the next step was to determine the most appropriate way of calculating it. If the objective of the *ex situ* conservation is to capture and hold the broadest genetic diversity possible of a species, the ideal definition of representativeness would be in genetic terms. Therefore, the formula to determine the genetic representation (GR) in percentage terms would be:

$$GR = (NCA * 100) / TNA$$

where TNA is the total number of alleles in the sum of all the loci studied in the target species within the spatial area (continent, country, region, etc.) of the collection to be evaluated, and NCA is the number of alleles of the loci of this species captured by this collection. This ideal determination of genetic representation entails a practical impediment. Knowing the total number of alleles that a species may have in a territory as large as a country (the usual size of a germplasm collection from a national program) or even much lower levels, is, in practice, an insurmountable task for any species (except for those that are known to be composed of very few populations). Given the context of plant

genetic resources for food and agriculture, this exception is almost nonexistent. Additionally, trying to calculate the GR leads indirectly to having represented 100% of the alleles if the sampling of all populations implies the germplasm collection. In other words, if calculating the GR of a germplasm collection involves collecting samples and germplasm from all the populations of the species within a work area, then, regardless of how difficult this task may be, the maximum representativeness would already be achieved as long as the appropriate criteria for intra-population representativeness have been followed for the collection.

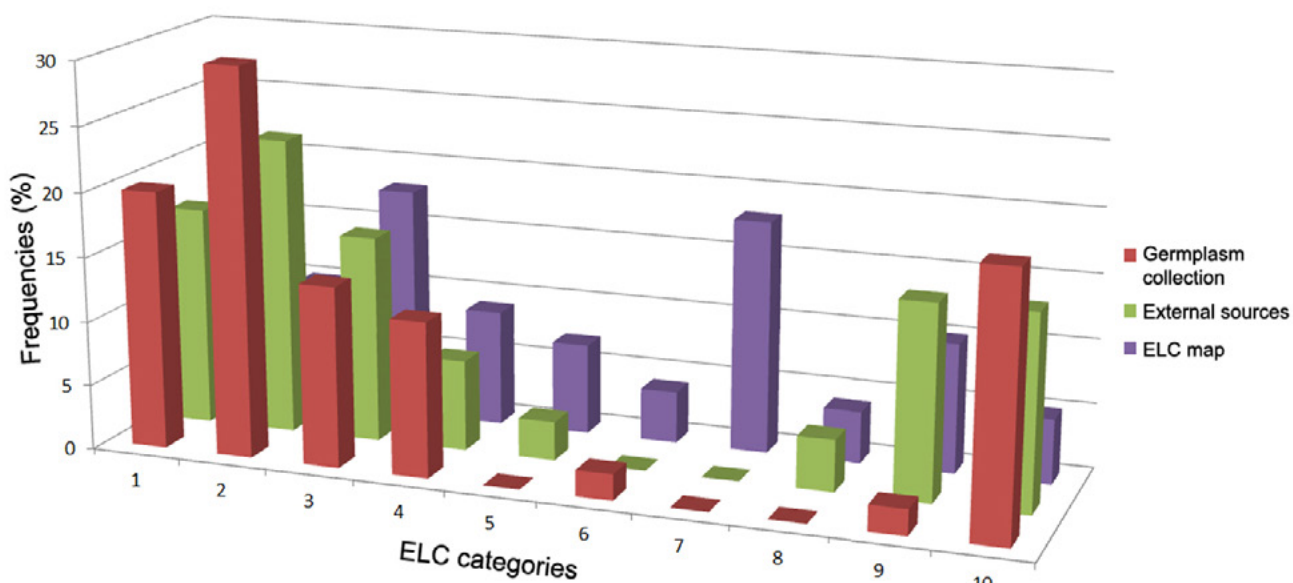
These practical and logistical difficulties have prompted the consideration of other alternatives to determine the representativeness of a collection. The issue of ecogeographic representativeness (ER) was raised by Parra-Quijano *et al.* (2008) in *ex situ* collections of crop wild relatives (CWR). The authors examined the possibility of using ecogeographic land characterization maps (such as those generated by ELCmaps) to find out how many environmental conditions in a given spatial work frame would be represented in a germplasm collection.



**Figure 33.** Comparison of the representation of each ELC category in the germplasm collection and the total availability of these categories in the ELC map, measured by frequency values (as a percentage).

As an example of this application, Fig. 33 shows a frequency distribution for each ELC category of a germplasm collection contrasted with the availability of these categories in the total spatial work frame. This fictitious example shows how the representativeness of a collection may, or may not, be biased according to the number of environmental units present in the work area (as is the case). The contrast between the values found in categories 2 and 7 reveals that the two distributions are highly dissimilar and that a Chi-squared test may determine an insignificant association between the two distributions. However, the most accurate determination of the ER is achieved using gap analysis. To do this, it is necessary to previously compile information from other sources external to the collection, such as other germplasm collections or any other data indicating the presence of populations of the target species (herbarium specimen sheets, botanical databases, bibliographic references, etc.). Then the frequency distribution of collection

sites for the collections being evaluated should be compared with that of external sources. This will enable a clear view of the environments that are underrepresented in the collection.



**Figure 34.** Comparison of the frequency distribution of collection sites in the target collection and presence of external sources about ten ELC categories. This also includes the distribution of the frequency of each category in the total of the ELC map.

Fig. 34 illustrates the previously mentioned comparison process. Using the same fictional data from the example in Fig. 33, this bar chart includes (in green) the frequency distribution of the ELC categories for presence data from external sources. In this case, the resemblance between distributions of the target collection and external sources is clear, and some differences are especially interesting. For categories 5 and 8, external sources indicate the presence of the species in that environmental unit, which is not represented in the collection. This shows that there are missing or empty ecogeographic data. These gaps may be useful for planning how to collect new germplasm, as one can prioritize visiting these environments since the location of these populations from external sources is known. It is important to clarify how presence data from external sources could be analysed. By taking the presence data provided by another germplasm collection as an external source, you can learn about the representativeness of the target collection globally; however, using these data to determine priority sites for collecting can lead to collect inter-collection duplicates.

### 8.3. Using Representa tool

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and Representa tool has been selected, the user should specify a series of parameters. Note that for this tool to work correctly, the user must have previously obtained an ELC map with ELCmaps tool, and the following product files of

this tool must be copied and pasted in the path CAPFITOGEN3/ELCmapas/ in *local mode*:

- mapa\_elc\_pais.grd
- mapa\_elc\_pais.gri
- Estadist\_ELC\_pais.txt

Where 'pais' is the name of the country or region for which the map was made. For *on server mode*, these elements must be uploaded through the button in front of each parameter, which allows selecting the file locally, and then uploading it to the server.

### **8.3.1 Initial parameters defined by the user**

#### **8.3.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### **8.3.1.2 Parameter: internet**

Explanation: Select this option if you want to use the Global Biodiversity Information Facility (GBIF) as an external source, connecting directly to its server, to transfer the information of the selected species to CAPFITOGEN3.

#### **8.3.1.3 Parameter: pasaporte**

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is called 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

#### **8.3.1.4 Parameter: geoqual**

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence

sites in terms of their georeferencing.

#### 8.3.1.5 Parameter: totalqual

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

#### 8.3.1.6 Parameter: fext

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you have input from outside sources (meaning any information source other than the target collection being analysed for representativeness) in the requisite format.

#### 8.3.1.7 Parameter: fuentex

Explanation: Applies only if 'fext' has been set as TRUE (✓ in *on server mode*). Please indicate the name of the file containing the input from external sources in the requisite format. If the file is called 'ExternalSources', then 'ExternalSource.txt' should appear in the field (because the table must be in tab-delimited text format). Please remember that this file should be saved in the folder called Pasaporte (CAPFITOGEN3/Pasaporte) for **local mode** or selected and uploaded to the server (through the button in front of the parameter).

#### 8.3.1.8 Parameter: geoqualfe

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if your table of externally sourced input was evaluated by GEOQUAL and you want to consider a minimum quality standard to be met by data to be included in the analysis. Information on the evaluation of the quality of the georeferencing should be available in the indicated columns.

#### 8.3.1.9 Parameter: totalqualfe

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). Determine the value of TOTALQUAL100 that includes the table of external sources to use as a threshold. Values from 0 (zero quality) to 100 (highest quality) are allowed.

#### 8.3.1.10 Parameter: duplibg

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you consider the input provided by other databanks or germplasm collections (external sources) to be gaps in the target germplasm bank (column TYPE-SOURCE with a value of 40 in the external sources table). Please note that if you select this option, you may create



collections of populations that are already represented in other collections, leading to duplicates between them. Caution: Setting `duplibg` as TRUE or ✓ when all the records in the external sources table come from passport banks (TYPESOURCE=40) will generate an error, since all of them will be discarded in the representativeness analysis.

#### 8.3.1.11 Parameter: `gbiffE`

Explanation: Applies only if 'internet' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to download externally sourced data from the Global Biodiversity Information Facility (GBIF) website. This option requires an Internet connection and is incompatible with the inclusion of externally sourced input provided by the user. If you select this option and also provide a table with externally sourced input, it will only take the latter into account.

#### 8.3.1.12 Parameter: `genero`

Explanation: Applies only if 'gbiffE' has been set as TRUE (✓ in *on server mode*). Type the name of the genus of the species to be analysed. This is the genus for which information will be downloaded from the GBIF website. Remember to capitalize the first letter.

#### 8.3.1.13 Parameter: `especie`

Explanation: Applies only if 'gbiffE' has been set as TRUE (✓ in *on server mode*). Type the name of the species (epithet only) to be analysed. This name will be placed next to the genus to request and download information from GBIF. If you wish to download information for the entire genus, type only an asterisk (\*). The epithet should be written entirely in lowercase.

#### 8.3.1.14 Parameter: `mapaelc`

Explanation: Enter the name of the file containing the ELC map generated by the ELCmapas tool. This map should be found in the ELCmapas folder, one of the folders making up the CAPFITOGEN directory (*local mode*). The map should be in DIVA-GIS format, made up of the two files with extensions '.grd' and '.gri', as generated by ELCmapas. In this text box, type the file name with the extension '.grd'. Thus, if the name of the map is 'mapa\_elc\_spain', enter 'mapa\_elc\_spain.grd'. For *on server mode*, select and upload the ELC map file on the server by clicking the button in front of the parameter when using the tool, or previously, saving the .grd and .gri files in the User's Files and Results area.

#### 8.3.1.15 Parameter: `statelc`

Explanation: Enter the name of the file with the table of the ELC map's descriptive statistics generated by the ELCmapas tool (the tool usually names this file 'Estadist\_ELC\_' plus the name of the country or region) for *local mode*. Like the ELC map, this file should also be located in the ELCmapas folder (CAPFITOGEN3/ELCmapas). Similarly, the name should be followed by the file extension, which in this case is '.txt' because the file is a table. Therefore, if the file is named 'Estadist\_ELC\_spain', it should be written 'Estadist\_ELC\_spain.txt'. For the on-server mode, select and

upload the .txt file on the server by clicking the button in front of the parameter when using the tool, or previously, saving the .txt file in the User's Files and Results area.

#### 8.3.1.16 Parameter: **distdup**

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites represent the same population. The value of zero (by default) excludes accessions with identical coordinates from the analysis of representativeness. The determination of the distance depends on biological (gene flow) and spatial (mean population sizes) conditions. This is a specific parameter for the target species, and it will often be necessary to consult an expert for his/her concept.

#### 8.3.1.17 Parameter: **resultados** (only for *local mode*)

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

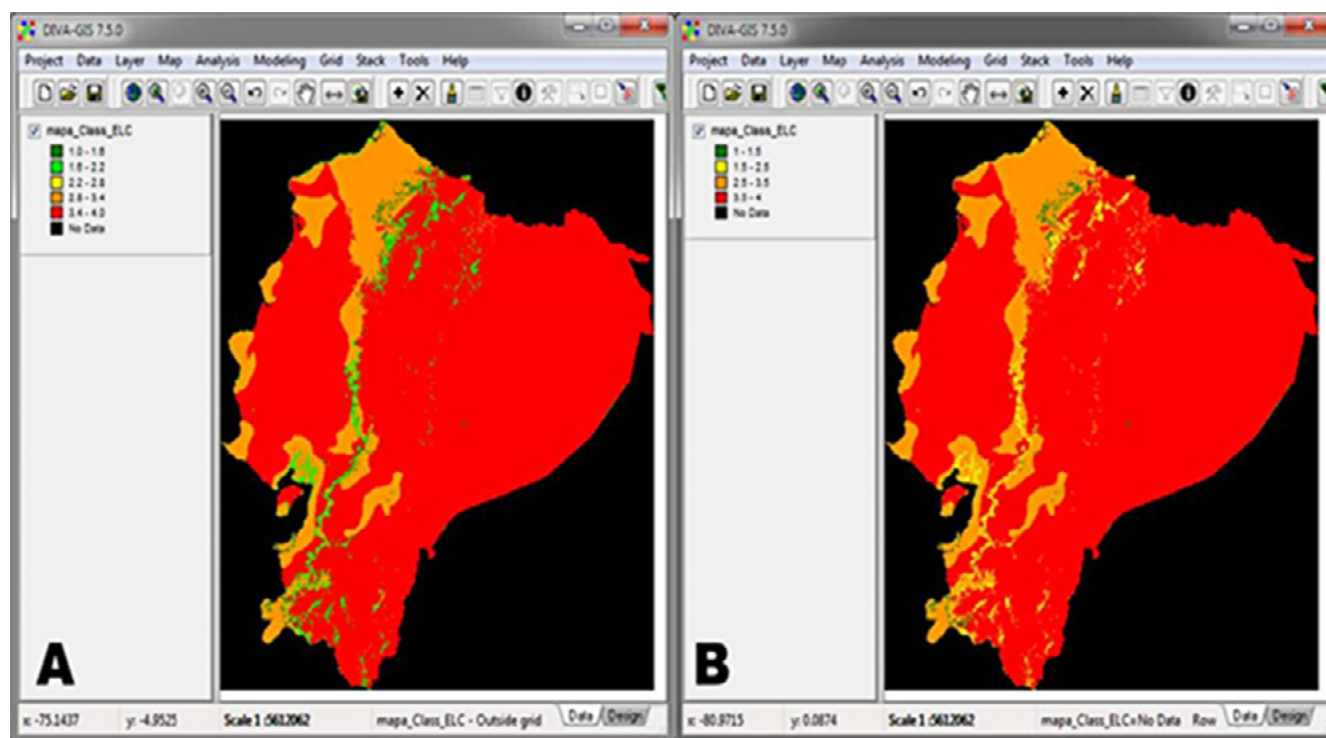
## 8.4. Results of Representa

In the path and folder created for 'resultados' (*local mode*) or in the User's Files and Results area (*on server mode*) up to five maps and up to five tables will be generated.

### 8.4.1 Maps

These are two vector point maps (shapefiles) and three raster maps (in .grid and .tif formats) that can be directly displayed in DIVA-GIS or other GIS software. If data from external sources are not introduced, there will be only three maps.

**8.4.1.1 'mapa\_Class\_ELC'.** This map divides the original ELC map categories into four groups according to their frequency across the whole territory. The frequency is divided based on quartiles. Group 1 corresponds to the lowest frequency (below the 0.25 quartile); group 2 is the medium-low frequency (between quartiles 0.25 and 0.5 or median); group 3 corresponds to medium-high frequency (between quartiles 0.5 or median, and 0.75), and group 4 corresponds to the highest frequency (above the 0.75 quartile). When this map in .grid format is opened in DIVA-GIS, five colours are displayed by default, so it is important to change the display to ensure that only four colours are seen. Each colour corresponds to a range representing each group value (1 to 4). This is illustrated in Fig. 35.



**Figure 35.** The appropriate visual configuration for Representa raster maps. A) View of the `mapa_Class_ELC.grd` file as opened in DIVA-GIS. B) Display adjusted to four colours (one per frequency group). The least frequent adaptive scenarios (low and medium-low frequency groups) appear in green and yellow.

**8.4.1.2 'mapa\_Class\_Sp'.** This map divides the original ELC map categories into four groups. These groups correspond to the division of ELC categories by frequency across the whole territory. The frequency is divided based on quartiles. Group 1 corresponds to the lowest frequency (below the 0.25 quartile); group 2 corresponds to medium-low frequency (between quartiles 0.25 and 0.5 or median); group 3 corresponds to medium-high frequency (between quartiles 0.5 or median, and 0.75), and group 4 corresponds to the highest frequency (above the 0.75 quartile).

**8.4.1.3 'mapa\_Tipo\_faltante'.** This map is another reclassification of the original ELC map categories. This map is only generated when the user enters data from external sources. This reclassification corresponds to criteria set out in the following table:

**Table 1.** Classification of ELC map categories according to priority criteria for future exploration.

Class	Difference between external sources and germplasm bank (DIF) <sup>1</sup>	Classification by frequency of species occurrence <sup>2</sup>	Classification by frequency of the category in the ELC map <sup>3</sup>
0	Not applicable	Not applicable	Not applicable
1	1	Low or medium-low	Low or medium-low
2	1	Low or medium-low	Medium-high or high
3	1	Medium-high or high	Low or medium-low
4	1	Medium-high or high	Medium-high or high
5	0.99-0.5	Low or medium-low	Low or medium-low
6	0.99-0.5	Low or medium-low	Medium-high or high
7	0.99-0.5	Medium-high or high	Low or medium-low
8	0.99-0.5	Medium-high or high	Medium-high or high
9	0.01-0.499	Low or medium-low	Low or medium-low
10	0.01-0.499	Low or medium-low	Medium-high or high
11	0.01-0.499	Medium-high or high	Low or medium-low
12	0.01-0.499	Medium-high or high	Medium-high or high
13	0 and NA	Not applicable	Not applicable

These classes are related to the priority level assigned to the visit or exploration of each ecogeographic category in a future collection. Class 1 comprises categories with the highest priority, while class 2 has a lower priority than class 1, and so on consecutively until class 13.

When the map is opened in DIVA-GIS it does not show the 13 classes with an individual colour for each class, but all 13 values into five colours. The correct display is achieved using DIVA-GIS to add eight more colours and adjusting the value ranges of each colour (as in previous maps) to the value of a class.

**1** This value is determined by comparing occurrences in external sources with those from germplasm collections/banks in each category according to the following formula:  $DIF = ES / (ES + GB)$  where ES refers to the number of occurrences from external sources while GB refers to the germplasm bank.

**2** This classification is the same as that shown in map 8.4.1.2.

**3** This classification is the same as that shown in map 8.4.1.1.

**8.4.1.4 'Shapefile\_Puntos\_BG.shp'.** Vector map (shapefile) representing the collection sites of the germplasm bank or the collection being evaluated for representativeness. The table that goes along with this map contains all fields of the FAO/Bioversity 2012 passport format.

**8.4.1.5 'Shapefile\_FE\_class.shp'.** Vector map (shapefile) representing the occurrences from external sources. The table accompanying this point map presents the following fields in addition to the data format from external sources:

**FE\_cat:** Category of the ELC map in which these are present.

**FE\_BG\_dif:** DIF value (see Table 1) for the ELC category in which these are present.

**Class\_Sp:** Indicates the quartile to which the category where the external source is present belongs, according to the species frequency.

**Class\_ELC:** Indicates the quartile to which the category where the external source is present belongs, according to the frequency of the same category in the ELC map.

**Tipo\_falt:** Indicates the class to which the category where the external source is present belongs, according to the classification given in Table 1.

## **8.4.2 Tables**

Just as with the maps, the list of tables may be reduced from five to four, depending on whether the user enters data from external sources or not.

**8.4.2.1 'Tabla Fuentes Externas clasificadas ExternalSourcesClassified'.** This, like the rest of the tables in CAP-FITOGEN3, is offered in tab-delimited text format with extensions .txt and .xls. This corresponds to the same table accompanying the shapefile in paragraph 8.4.1.5 and contains the same variables.

**8.4.2.2 'Tabla Resultados Representatividad RepresentativenessResults'.** This table shows the results of the representativeness evaluation, whether data from external sources has been included or not. With this table, it is possible to create bar graphs in Excel as shown in Figs. 33 and 34. Finally, this table shows all the information required to calculate the parameters in Table 1, including the class value used to define priorities.

**8.4.2.3 'Tabla Resultados X2 Results'.** This table shows the results of the Chi-squared test to determine the degree of association between two distributions. If data from external sources have been introduced, this table will contain two Chi-squared test results: distribution bank/collection (or GB) vs. external sources (ES), and bank/collection vs. distribution of total frequencies of the ELC map categories.

**8.4.2.4 'TablaClasificacionCuartilesEspecie\_QuartileClassificationSpecies' and 'TablaClasificacionCuartilesMapaELC\_QuartileClassificationELCmap'.** These two tables show values of the quartiles 0.25, 0.5 (median) and 0.75 for the distribution of species frequencies and ELC map categories.

**8.4.2.5 'Genbank\_ELC'.** This table corresponds to the same information entered in the passport format plus column 'BGcat' where you can see the category of the ELC map that corresponds to each entry according to its location.

## 8.5. References

Brown, A.H.D. 1989. The case for core collections. In: Brown, A.H.D., Frankel, O.H., Marshall, D.R., Williams, J.T. (eds.) The use of plant genetic resources. Cambridge University Press, Cambridge, UK.

Crossa, J., Vencovsky, R. 1994. Implications of the variance effective population size on the genetic conservation of monoecious species. *Theoretical and Applied Genetics* 89:936–942

Crossa, J., Vencovsky, R. 1997. Variance effective population size for two-stage sampling of monoecious species. *Crop Science* 37:14–26

Crossa, J., Vencovsky, R. 2011 Chapter 5: Basic sampling strategies: theory and practice. In: Guarino, L., Ramanatha Rao, V., Goldberg, E. (eds.) *Collecting Plant Genetic Diversity: Technical Guidelines – 2011 Update*. Bioversity International. Available online (accessed 6 November 2013) [http://cropgenebank.sgrp.cgiar.org/index.php?option=com\\_content&view=article&id=671](http://cropgenebank.sgrp.cgiar.org/index.php?option=com_content&view=article&id=671)

Parra-Quijano, M.; Draper, D.; Torres, E., Iriondo, J.M. 2008. Ecogeographical representativeness in crop wild relative ex situ collections. p. 249-273. In Maxted, N., Ford-Lloyd, B.V., Kell, S.P., Iriondo, J.M., Dulloo, M.E., Turok, J. (eds.) *Crop wild relative conservation and use*. CAB International, Wallingford.

Yonezawa, K., Nomura, T., Morishima, H. 1995. Sampling strategies for use in stratified germplasm collections. p. 35-53. In: Hodgkin, T., Brown, A.H.D., van Hintum, Th.J.L., Morales, E.A.V. (eds.) *Core collections of plant genetic resources*. John Willey & sons, Chichester, UK.





Regional Workshop, Bogotá (Colombia), March 2013.





# 9 | DIVmapas Tool

## 9.1. Spatial representation of local diversity

In 2012, a study was published on the presentation of spatial patterns of genetic diversity from neutral markers of the microsatellite type in the case of *Annona cherimola* (van Zonneveld *et al.*, 2012). The study aims to show a different way of displaying the distribution of genotypic diversity, based on the estimation of parameters belonging to population genetics. However, in this case, before they are applied to all samples at once, diversity is estimated at a local level with the determination of neighbourhoods or areas of influence. The results of putting together all the results from each neighbourhood led to a map that clearly shows where the diversity ‘hot spots’ are located. The application of this methodology to the *ex situ* and *in situ* conservation of plant genetic resources is evident.

This is not the first GIS or geostatistical approach used to analyse genetic diversity, as there have also been earlier interpolations of genetic data (Hoffman *et al.*, 2003). However, the methodology used by van Zonneveld and his collaborators is very practical and simple in terms of its analysis and interpretation.

Later, Thomas *et al.* (2012) applied the same methodology to 993 individuals characterized by cocoa (*Theobroma cacao*) microsatellites, in addition to other analyses, to identify evolutionary processes in this cultivated plant.

Based on the publication of these developments, it became possible to understand the steps involved in the process of obtaining a map of this type. The methodology could be replicated as the only element that varies is the genetic parameter that is calculated from the samples making up a neighbourhood. Thus, if the parameter expresses the genetic differences between samples from a specific neighbourhood, the map could be called a ‘diversity map’. The “DIVmapas” tool was developed based on this methodology, and its application broadened beyond genotypic characterization data.

It is very important to note that these maps show genotypic diversity at the intraspecific level, one aspect that differentiates them notably from maps showing the richness of species or phylogenetic diversity maps, which work at the inter-specific level.

Illustrating diversity in the form of maps has multiple advantages over the ways in which these results are usually presented. Diversity maps, based on the original version developed by van Zonneveld *et al.* (2012), can simply and quickly identify those areas or regions with a high concentration of variability. This type of map becomes a powerful tool for decision-making regarding *ex-situ* and *in situ* conservation.

### **9.1.1 Why a map of ecogeographic diversity?**

The ecogeographic diversity of a cluster of accessions is one way of measuring the differences occurring between the adaptive scenarios where these accessions are sourced, or in other words, the collection sites. The term ‘adaptive scenario’ is used rather than ‘environment’, since only the abiotic environmental features with the greatest influence on the distribution and occurrence of the target species are considered when calculating ecogeographic diversity, as opposed to using all the environmental characteristics available.

Ecogeographic diversity, like any other kind of diversity, is determined based on germplasm characterization data. Ecogeographic characterization is carried out by extracting information for each coordinate using GIS software, which has been previously loaded with layers of environmental information.

The display of ecogeographic diversity as a map similar to those developed by van Zonneveld *et al.* (2012) facilitates the comparison between areas or regions based on the difference between the adaptive scenarios where the accessions occur. The zones or regions where the greatest differences occur can be translated directly into zones where one may expect to find germplasm with more divergent adaptations. This may also indirectly indicate the possible occurrence of greater genotypic or phenotypic diversity. The determination of areas with greater genotypic or phenotypic diversity is best when carried out using genotypic and phenotypic characterization data, respectively. However, in the absence of these data, a map of ecogeographic diversity may serve as an interim solution while the accessions are characterized in genotypic and/or phenotypic terms. In any case, the ideal setting for diversity analysis under this new methodology is when maps may be obtained for the three types of characterization, as the contrast offers a very complete biological view of the status of plant genetic resources occurring within a work frame.

## 9.2. Procedure for obtaining diversity maps using DIVmapas tool

DIVmapas is an application developed based on the application created by van Zonneveld *et al.* (2012) for the custard apple (*Annona cherimola*). However, it has some differences from the original methodology that become very clear when comparing the two processes. This section will show, step by step, how DIVmapas tool creates diversity maps.

DIVmapas tool determines ways of measuring local diversity. For instance, it compares accessions collected in a grid-shaped area of a certain size with other neighbourhoods (zone of influence), using ecogeographic, phenotypic, or genotypic input. Note that from this point on we shall be referring to accessions rather than samples, as the tool is intended to be used in the field of plant genetic resources. However, this does not imply that the tool cannot be used in other biological fields. As a result, DIVmapas tool offers a graphic illustration that reflects the values of the diversity measurements in a map, which helps to visualize genetic diversity hot spots.

It is important to note that DIVmapas tool, like other tools included in this manual and many other GIS and ecogeographic tools for plant genetic resources, requires each accession to be properly geo-referenced. Chapter 3 of this manual refers to GEOQUAL tool, which provides information on the quality of the georeferencing of the germplasm collection site. It is advisable to use this tool before using DIVmapas tool so that only accessions with sufficiently high georeferencing quality are considered when obtaining diversity maps. In any case, accessions without coordinates (fields DECLATITUDE or LATITUDE and DECLONGITUDE or LONGITUDE) will not be included in the analysis performed by DIVmapas tool.

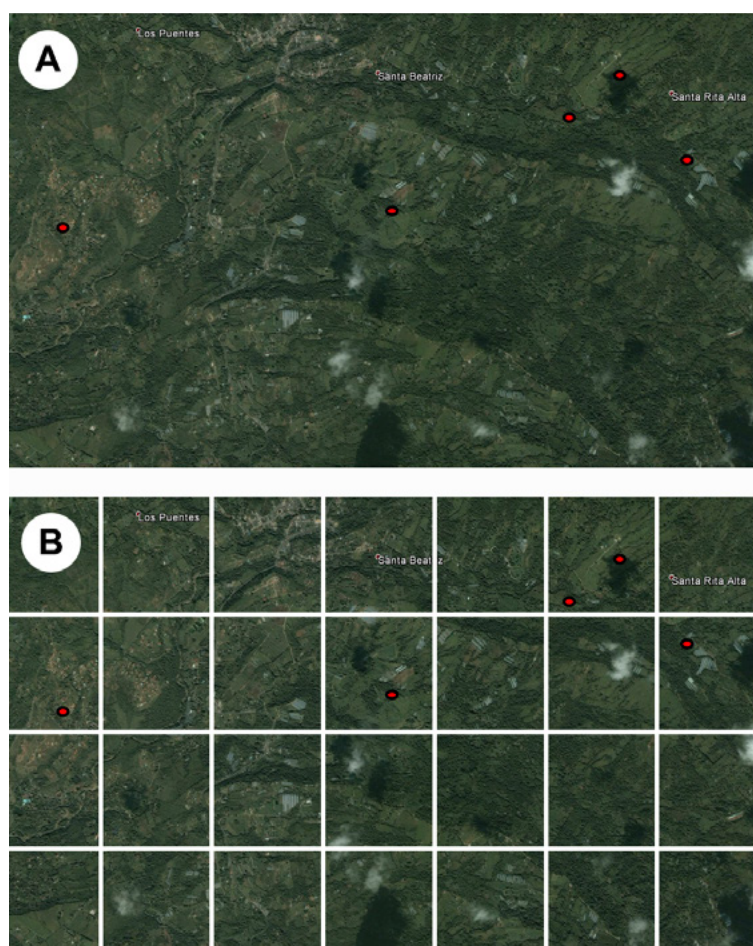
The second important point is that if you need to obtain phenotypic or genotypic diversity maps, details of the characterization of each type must be arranged according to the format usually supplied in the 'Formatos'/'Formats' folder (Excel .xls files). If you require an ecogeographic diversity map, please note that DIVmapas tool includes the

same germplasm ecogeographic classification process as ECOGEO tool (Chapter 5). Therefore, it is not necessary to prepare characterization data tables or matrices; simply indicate the ecogeographic variables that you wish to use to characterize the accessions.

DIVmapas tool will take advantage of all the valid characterization information available and, accordingly, it will create diversity maps for each aspect. Thus, the list of accessions characterized on a genotypic, phenotypic, or ecogeographic basis may either match (which facilitates the interpretation of results) or not. Identification codes for the accessions in the genotypic or phenotypic characterization tables must be included in the FAO/Bioversity 2012 passport table containing geo-referencing information from the collection sites.

Once these conditions are clear, the following points show how DIVmapas tool generates diversity maps, regardless of the characterization data used for this purpose.

### **9.2.1. Distribution of collection sites and generation of grid**

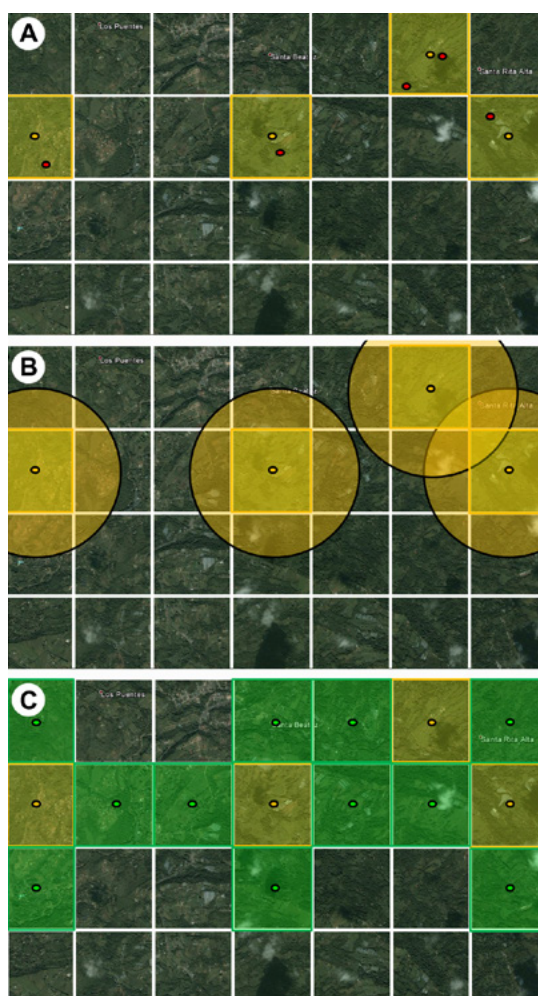


**Figure 36.** First step. A) spatial distribution of the collection sites and B) overlay of cell dimension grid (resolution) selected by the user.

A workspace (x-min, y-min, x-max, and y-max where x is latitude and y longitude) is generated using the coordinates for each collection site. A square grid or set of cells defined by the user is then overlaid (see Fig. 36). Additionally, a layer including the centroids of each cell in the grid is loaded. Each centroid has an identification code.

### **9.2.2. Selection of cells with accessions and neighbourhood cells**

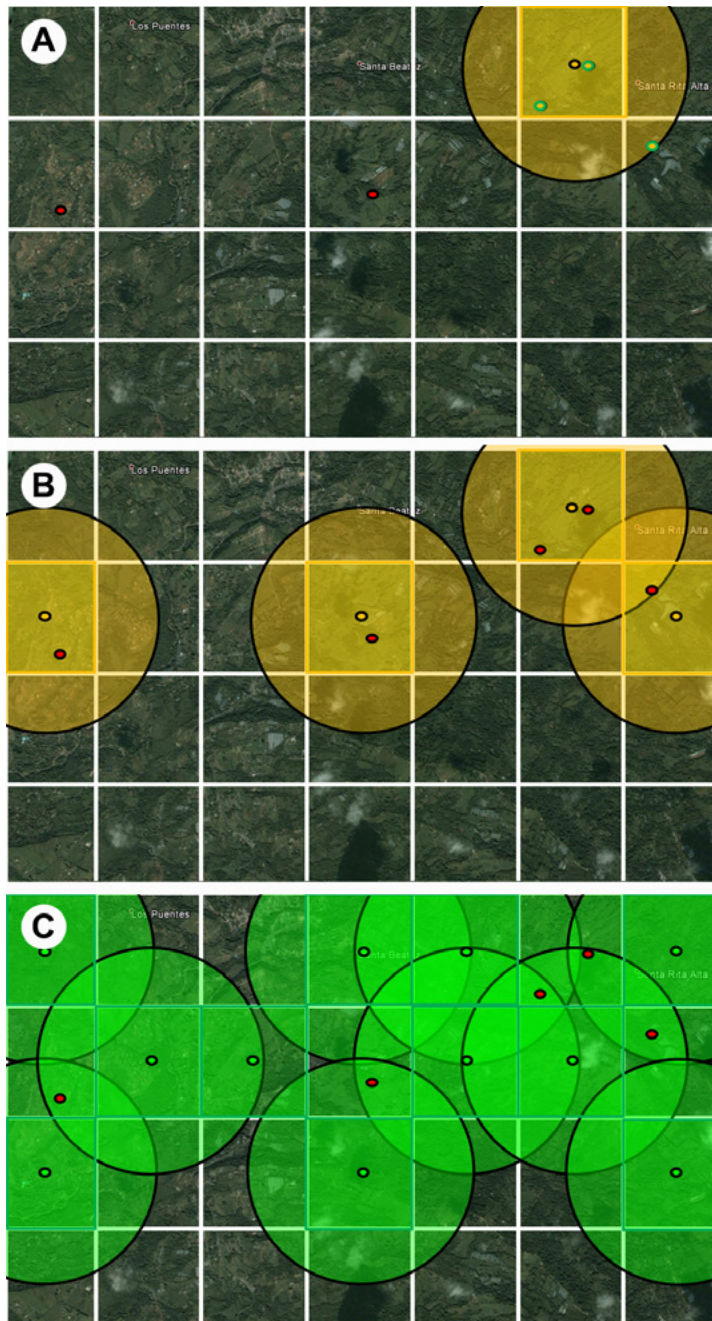
The cells with accessions are selected from the total number of cells making up the grid. The user also determines an area of influence by indicating the radius of a circular area. This is related to the reproductive biology of the species, its gene flow as well as any handling and dispersal of human origin, particularly if this is a cultivated form. This area of influence is used to determine the neighbourhood cells, which are cells without accessions lying close to those initially selected (cells with accessions). For a cell to qualify as a neighbourhood cell, its centroid should fall within the projection of the circular area of influence drawn from the centroid of each cell containing accessions. The process to select cells with accessions and neighbourhood cells is shown in Fig. 37.



**Figure 37.** Second step. A) Determination of cells with accessions and their centroids, B) projection of the areas of influence from the centroids of cells with accessions, and C) determination of neighbourhood cells.

### **9.2.3. Determination of accessions linked to cells with accessions and neighbourhood cells**

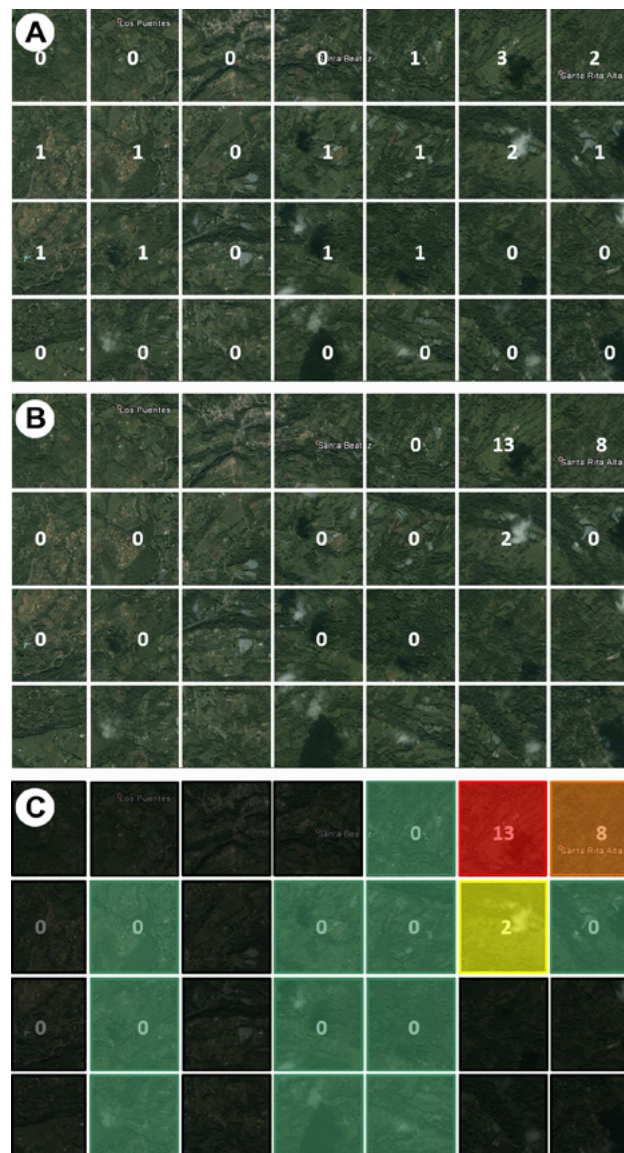
The circular areas of influence are again projected from the centroids of the cells with accessions and the neighbourhood cells. The ensuing list of accessions falling into each area is assigned an identification code for its respective centroid (see Fig. 38).



**Figure 38.** Third step. A) Determination of accessions occurring within the area of influence of a single cell, B) determination of accessions occurring within the areas of influence of cells with accessions, and C) determination of accessions occurring within the areas of influence of neighbourhood cells.

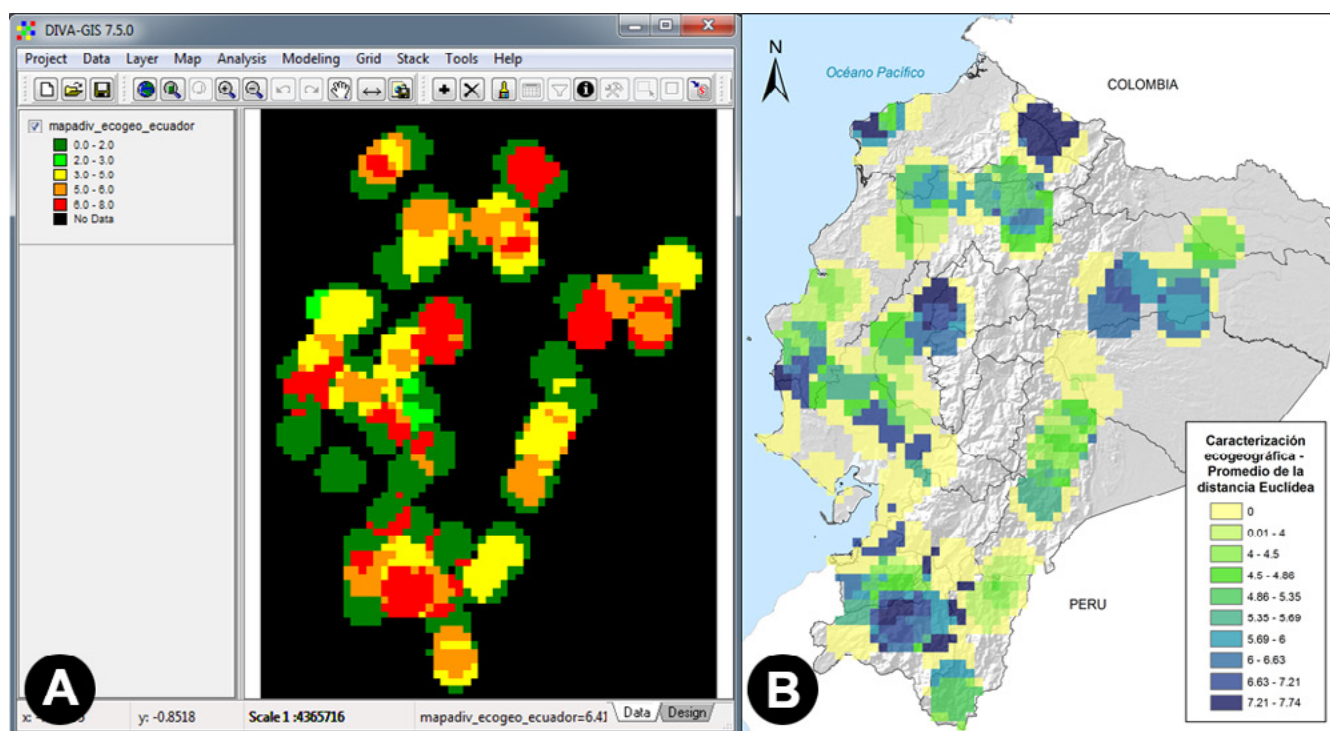
### 9.2.4. Obtaining final diversity maps

The list of accessions per cell may be used to obtain the initial matrices where the phenotypic, genotypic, or ecogeographic characterization data (depending on the data entered by the user) appear in columns and the accessions for each centroid are identified by their ACCENUMB value in rows. Thus, determining the number of cells with accessions and neighbourhood cells indicates the number of initial matrices to be obtained. The process to standardize data is applied to each initial matrix when the data involves quantitative variables. Subsequently, a distance or similarity/dissimilarity coefficient is applied, which also produces a diagonal distance matrix. The average distance of the accessions included is calculated based on this matrix, and this value is assigned to each centroid code and its respective cell. This allows R to produce raster cell maps reflecting the values assigned (see Fig. 39).



**Figure 39.** Fourth step. A) Number of accessions analysed by cell, B) values assigned to cells of an average genotypic, phenotypic, or ecogeographic distance, and C) assignment of colours graded according to the average values of distance.

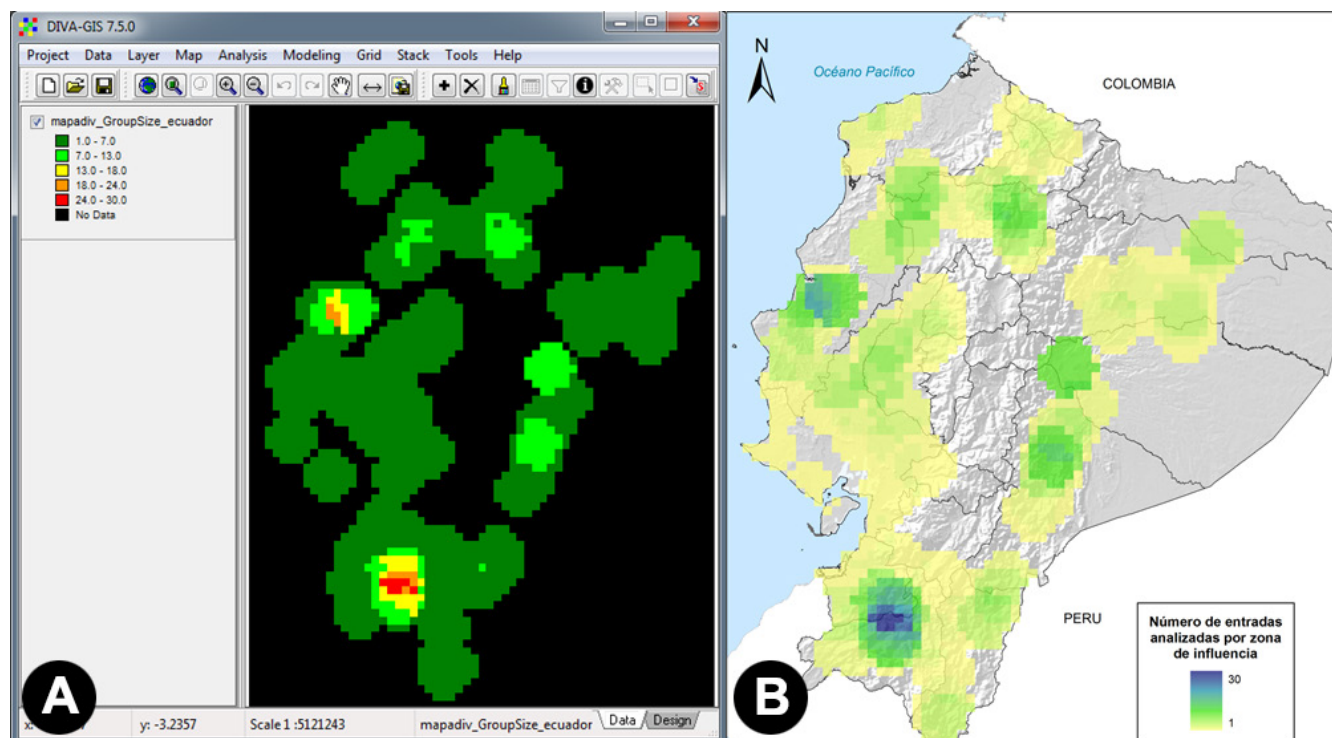
In the case of genotypic characterization, in addition to the average distance or dissimilarity, other genetic parameters may be calculated, such as Nei's measure of genetic diversity (1987), or the proportion of polymorphic markers for each group of accessions within each area of influence. R calculates these parameters using the initial characterization matrices. Finally, when the raster file of the cells whose values were assigned by the diversity parameters is displayed in DIVA-GIS, the software assigns each one a specific colour from a graded colour swatch. This allows you to quickly observe the locations with the highest levels of diversity, as measured by the average values of distance/dissimilarity or by other genetic parameters (see Fig. 40).



**Figure 40.** Display of diversity maps. A) A map of ecogeographic diversity for the Ecuadorian peanut (*Arachis hypogaea*) collection generated by DIVmapas tool opened in DIVA-GIS, and B) the same map in its final version.

If different kinds of characterization data have been entered, several maps will appear as follows: one for ecogeographic characterization, one for phenotypic characterization, and one or more for genotypic characterization. This depends on whether the user has requested the calculation of one or more parameters. A map of the number of accessions analysed by cell is also generated, as shown in Fig. 41, corresponding to Fig. 39 part A. This last map can be used as a support to determine whether there is any potential bias in the collection or interpretation of the patterns found in the diversity maps.





**Figure 41.** View of the map with the number of accessions analysed by cell. A) Map opened in the DIVA-GIS program, and B) the final version of the same map.

### 9.2.5. Other analyses

DIVmapas tool also permits other types of analysis to be performed, particularly when characterization data of different types have been entered. DIVmapas thus asks the user if he/she wants to perform a cluster analysis or a management analysis in the same way as ECOGEO tool. The user may also request Mantel test (1967) comparisons between distance matrices for all accessions. DIVmapas automatically creates a matrix of geographical distances between all the collection sites and enters this matrix into the paired matrix correlations.

## 9.3. Formats for data entered into DIVmapas

To ensure that DIVmapas tool works properly, enter the different kinds of information in the indicated formats. These formats are usually located in the 'Formatos' folder within the CAPFITOGEN structure of folders and files. Inside this folder, you will find another one with the name 'Formatos DIVmapas', and within this last folder, you will find five Excel files.

### **9.3.1. Model of passport data**

As in other CAPFITOGEN tools, the passport data must be entered using the FAO/Bioversity 2012 format with minor modifications (file 'ModeloDatosPasaporte\_FAO\_BIOVERSITY\_2012.xls' available to users at <https://drive.google.com/drive/folders/1xCnllZgzW0uDeClDvcxbADv9H583xzpn?usp=sharing>). This format is described in the chapter about TesTable tool. Since DIVmapas allows you to work with data evaluated based on its geo-referencing quality using GEOQUAL, in addition to the normal passport data model, there is also a model with additional fields for the GEOEQUAL evaluation values. However, the easiest way to use GEOEQUAL-evaluated passport data is to directly load the passport table containing all the GEOEQUAL evaluation data called 'PasaporteOriginalEvaluadoGEOQUAL.txt'. Remember that this table must be in a tab-delimited text file format and must be exported from Excel in this format. It should be saved in the 'Pasaporte' folder in the CAPFITOGEN structure of folders and files.

### **9.3.2. Phenotypic data model**

When the contents of the phenotypic data format ('ModeloDatosFenotipicos.xlsx' file) are displayed, a green column called 'ACCENUMB' will appear (which must be filled in). This corresponds to the same ACCENUMB code used for the passport data table. The order in which the codes are given is not relevant. Since phenotypic characterization data is not always available for all the accessions in the passport table, the number of accessions in the phenotypic data table may be lower than the number of passport data. What should not happen is for accessions or ACCENUMB codes to appear in the phenotypic data table but not in the passport data table. This will generate a processing error.

The other columns in this format are named 'D1', 'D2' and 'D3'. These names represent the names of phenotypic descriptors 1, 2, and 3. The format only includes three descriptor columns; however, in theory, there can be as many descriptors as the user makes available, extending the sequence from 'D4' to as many as necessary. Their names may be changed (e.g., 'D1' to 'NGRANOS') for greater ease of use. Should you wish to change the names, there are three recommendations to remember. First, there must be no spaces in the name. Secondly, the name must include at least eleven characters. Thirdly, no name must be repeated. The third condition may generate an error.

The coding of the phenotypic variables imposes certain conditions. Variables, whether quantitative or categorical, must be expressed numerically. For categorical variables, the names of the states written with alphabetic or non-alphabetic characters when they were characterized must be changed to numeric codes, with no dashes, periods, commas, or spaces. Any missing data should be coded as 'NA'.

Finally, please note that DIVmapas tool only recognizes information in tables when it is in tab-delimited text format. As a result, once the phenotypic data has been completed in Excel according to the previously mentioned requirements, the table must be exported in tab-delimited text format and saved in the 'Pasaporte' folder together with the other characterization data tables and the passport data table.

### **9.3.3. Model of the table of types of phenotypic variables**

If you wish to use available phenotypic characterization data to generate a diversity map using DIVmapas, in addition to providing the phenotypic data table given in 9.3.2, you must also fill in the table called 'ModeloTablaNaturalezaVariables.xlsx'. This table indicates the nature of each phenotypic variable or descriptor included in the phenotypic data table. This Excel file contains two worksheets. The first ('Natvariables') is the phenotypic variables type table, which contains only three columns. In the first column, named 'ID', a number is assigned to each variable in consecutive form (1, 2, 3...) so that each row in the table corresponds to a phenotypic variable or descriptor in the phenotypic data table. The second column, named 'NOMVAR', corresponds exactly to the names assigned to the variables or descriptors in the phenotypic data table. The third and last column is named 'NATVAR'; it indicates the nature of the variable or corresponding descriptor. When you place the cursor over a cell, the list of possible values for this column appears, namely: binary symmetric, binary asymmetric, nominal, ordinal, or quantitative.

Finally, the 'Observations' worksheet contains some guidelines and tips to help users to fill in the 'Natvariables' spreadsheet.

At the end of the process, export the table with the nature of variables using tab-delimited text format and save it in the 'Pasaporte' folder in the same way as the other input data tables.

### **9.3.4. Genotypic data model**

As mentioned above, DIVmapas is a way of creating diversity maps based on genotypic germplasm characterization by analysing information from molecular markers as if these were of the dominant type. This means that the genotypic data table (in the Excel file 'ModeloDatosGenotipicos0\_1.xlsx') contains absence/presence variables that are encoded as 0 and 1, respectively. As the structure of this table is very similar to the phenotypic data table, it should be completed in the same way, except that all the variables or descriptors in the genotypic data table correspond to asymmetric binary variables and, thus, must be encoded with values 0 and 1.

As with the phenotypic information, DIVmapas tool only recognizes information in tables when it is in tab-delimited text format. Accordingly, once the data has been completed in Excel as indicated, the table must be exported in tab-delimited text format and saved in the 'Pasaporte' folder together with the other characterization data tables and the passport data table.

## **9.4. Using DIVmapas tool**

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and DIVmapas tool has been selected, a series of parameters must be specified by the user.

### **9.4.1. Initial parameters defined by the user**

#### **9.4.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### **9.4.1.2 Parameter: pais**

Explanation: Select the country/region where all or most of the data accessions you wish to analyse were collected. If accessions have been collected from more than one country, you may select a region, subcontinent, or continent. You can also use rLayer tool to produce your own work frames and select them here.

#### **9.4.1.3 Parameter: pasaporte**

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

#### **9.4.1.4 Parameter: geoqual**

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### **9.4.1.5 Parameter: totalqual**

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

#### 9.4.1.6 Parameter: **buffy**

Explanation: Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish ecogeographic information to be extracted from a circular area around the collection site. Leaving this option unchecked means that information is extracted only from the point indicated by the collection site coordinates. This option is very useful, for example, when most of the collection sites correspond to markets and not directly to the growers' farms, or when the actual location of the collection sites is uncertain even though GEOQUAL can grant high georeferencing quality. This feature is explained in the chapter dedicated to ECOGEO tool.

#### 9.4.1.7 Parameter: **tamp**

Explanation: Applies only if 'buffy' has been set as TRUE (✓ in *on server mode*). Specify the radius (in km) of a circular area around the point indicated by the collection site coordinates from which the ecogeographic information is to be extracted. The values extracted from the circular area will be averaged to obtain a single value. The information will be extracted from those cells whose centroid is within the circular area.

#### 9.4.1.8 Parameter: **ecogeo**

Explanation: Indicate here (with the TRUE in *local mode* or ✓ in *on server mode*) if you are interested in carrying out an ecogeographic characterization of the germplasm collection/observation sites.

#### 9.4.1.9 Parameter: **resol1**

Explanation: Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution or precision regarding some coordinates (points X and Y) but requires greater computing capacity than 5x5 km. Resolutions of 10x10 and 20x20 km may only be used for large countries, subcontinents, or continents.

#### 9.4.1.10 Parameter: **bioclimsn**

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to use bioclimatic variables (annual or monthly average, maximum or minimum temperatures, monthly or annual rainfall, vapor pressure, etc.) to calculate ecogeographic diversity.

Explanation: Applies only if 'bioclimsn' has been set as TRUE (✓ in *on server mode*). List (*local mode*) or select (*on server mode*) the bioclimatic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of bioclimatic variables'); copy the names and paste them separated by a semicolon (;). You will also find a blocked line with the 19 bioclim variables ready to use; simply remove the initial # symbol. In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will

appear in the box in front of parameter 'bioclimv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 9.4.1.12 Parameter: edaphsn

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to use edaphic variables (texture, depth, pH, etc.) to calculate ecogeographic diversity.

#### 9.4.1.13 Parameter: edaphv

Explanation: Applies only if 'edaphsn' has been set as TRUE (✓ in *on server mode*). List (*local mode*) or select (*on server mode*) the edaphic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of edaphic variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'edaphv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 9.4.1.14 Parameter: geophysn

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to use geophysical variables (related to terrain and sunlight) to calculate ecogeographical diversity.

#### 9.4.1.15 Parameter: geophysv

Explanation: Applies only if 'geophysn' has been set as TRUE (✓ in *on server mode*). List (*local mode*) or select (*on server mode*) the geophysical variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of geophysical variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'geophysv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 9.4.1.16 Parameter: latitud

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) to include the latitude (Y) as a geophysical variable to be analysed.

#### 9.4.1.17 Parameter: longitud

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) to include the longitude (X) as a geophysical variable to be analysed.

#### 9.4.1.18 Parameter: phenotip

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a phenotypic diversity map. For this, you must have access to data for phenotypic characterization or evaluation (morphology, phenology, productivity, resistance, etc.) in the specified format. Please remember to include the name of the extension. For example, if the table is called 'phenotypes', in this space you must write 'phenotypes.txt'. Remember that this table must be located in the 'Pasaporte' folder of CAPFITOGEN3 *local mode* tools data structure. For *on server mode*, the table will be uploaded at the moment of configuring the tool or before this process so that it can be available in the User's Files and Results area.

#### 9.4.1.19 Parameter: phenot

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Indicate the name of the text file containing the data from the phenotypic characterization in the format indicated in *local mode*. Please remember to include the name of the extension. For example, if the table is called 'phenotypes', in this space you must write 'phenotypes.txt'. For *on server mode*, you must click on the button that allows you to search, select, and upload the corresponding file.

#### 9.4.1.20 Parameter: phenotv

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Indicate the name of the text file containing the table describing the nature of each phenotypic variable in the format indicated for *local mode*. Please remember to include the name of the extension. For example, if the table is called 'phenotypevariables', in this space you must write 'phenotypevariables.txt'. This table must describe all the variables included in the table with the characterization data (see the previous parameter). For *on server mode*, you must click on the button that allows you to search, select, and upload the corresponding file.

#### 9.4.1.21 Parameter: genotip

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a genotypic diversity map. For this, you must have access to data for phenotypic characterization or evaluation (such as the presence or absence of markers as zero and one) in the specified format. Remember that this table must be located in the 'Pasaporte' folder (*local mode*) of the CAPFITOGEN3 tools data structure. For *on server mode*, the table will be uploaded at the moment of configuring the tool or before this process so that it can be available in the User's Files and Results area.

**9.4.1.22 Parameter: *genot***

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Indicate the name of the text file containing the genotypic characterization data in the format indicated for *local mode*. Please remember to include the name of the extension. For example, if the table is called 'genotypes', in this space you must write 'genotypes.txt'. For *on server mode*, you must click on the button that allows you to search, select, and upload the corresponding file.

**9.4.1.23 Parameter: *neigd***

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Select this option if you wish to obtain a map of Nei's average index of genetic diversity (1987), a map of the average proportion of polymorphic markers, and a map of the number of accessions analysed by cell.

**9.4.1.24 Parameter: *csimilar***

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Indicate the similarity coefficient that you want to use to generate the map of average genotypic distance. 1 = Jaccard Index (1901), 2 = Simple Matching Coefficient (SMC) Sokal & Michener (1958), 3 = Sokal & Sneath (1963) (S5 coefficient of Gower & Legendre), 4 = Rogers & Tanimoto (1960), 5 = Dice (1945), 6 = Hamann coefficient, 7 = Ochiai (1957), 8 = Sokal & Sneath (1963) (S13 coefficient of Gower & Legendre), 9 = Pearson Phi coefficient, 10 = S2 coefficient of Gower & Legendre. Distance (d) is obtained as  $d = \sqrt{1-s}$  where s is the similarity coefficient.

**9.4.1.25 Parameter: *rgrid***

Explanation: Choose the cell size (in km) for the diversity map/maps to be generated. This parameter is restricted to the following values: 1, 5, 10, 50, and 100 km (if you choose another value, this will produce an error).

**9.4.1.26 Parameter: *buffer***

Explanation: Choose the radius of the circular area of influence or neighbourhood (in km). This area is created based on each cell centroid on the map showing collection sites and generates clusters using accessions whose collection sites are included. The value of the indexes and average distances of each cluster will be assigned to the cell from whose centroid the area of influence was drawn.

**9.4.1.27 Parameter: *ecogeoclus***

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a cluster analysis for accessions with an ecogeographic characterization.



#### 9.4.1.28 Parameter: **ecogeoclustype**

Explanation: Applies only if 'ecogeoclus' has been set as TRUE (✓ in *on server mode*). Choose the type of hierarchical cluster to be used for ecogeographic clusters: 'single' = nearest neighbour, 'complete' = more compact neighbourhood, 'ward' = method of minimum variance of Ward, 'mcquitty' = McQuitty's method, 'average' = average similarity (UPGMA), 'median' = similarity of the median, 'centroid' = geometric centroid, 'flexible' = flexible Beta.

#### 9.4.1.29 Parameter: **ecogeopca**

Explanation: Applies only if 'ecogeo' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to perform an analysis of principal components for accessions with an ecogeographic characterization.

#### 9.4.1.30 Parameter: **ecogeopcaxe**

Explanation: Applies only if 'ecogeopca' has been set as TRUE (✓ in *on server mode*). Indicate here the number of components to retain within the PCA analysis. This number should always be lower than the number of ecogeographic variables. This means that only the 'retained' principal components will be shown in the results tables.

#### 9.4.1.31 Parameter: **phenoclus**

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to obtain a cluster analysis of all accessions that include phenotypic information.

#### 9.4.1.32 Parameter: **phenoclustype**

Explanation: Applies only if 'phenoclus' has been set as TRUE (✓ in *on server mode*). Choose the type of hierarchical cluster to be used for phenotypic clusters: 'single' = nearest neighbour, 'complete' = more compact neighbourhood, 'ward' = method of minimum variance of Ward, 'mcquitty' = McQuitty's method, 'average' = average similarity (UPGMA), 'median' = similarity of the median, 'centroid' = geometric centroid, 'flexible' = flexible Beta.

#### 9.4.1.33 Parameter: **phenopca**

Explanation: Applies only if 'phenotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform a principal component/coordinate analysis of all accessions that include phenotypic information.

#### 9.4.1.34 Parameter: **phenopcaxe**

Explanation: Applies only if 'phenopca' has been set as TRUE (✓ in *on server mode*). Indicate here the number of com-

ponents/coordinates to retain within the PCA/PCoA analysis. This number should always be lower than the number of phenotypic variables. This means that only the 'retained' principal components will be shown in the results tables.

#### 9.4.1.35 Parameter: phenovarq

Explanation: Applies only if 'phenopca' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if all the phenotypic variables/descriptors correspond to quantitative variables.

#### 9.4.1.36 Parameter: genoclus

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform a cluster analysis of all accessions that include genotypic information.

#### 9.4.1.37 Parameter: genoclustype

Explanation: Applies only if 'genoclus' has been set as TRUE (✓ in *on server mode*). Choose the type of hierarchical cluster to be used for genotypic clusters: 'single' = nearest neighbour, 'complete' = more compact neighbourhood, 'ward' = method of minimum variance of Ward, 'mcquitty' = McQuitty's method, 'average' = average similarity (UPG-MA), 'median' = similarity of the median, 'centroid' = geometric centroid, 'flexible' = flexible Beta

#### 9.4.1.38 Parameter: genopco

Explanation: Applies only if 'genotip' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform an analysis of the principal coordinates of all the accessions that include genotypic information.

#### 9.4.1.39 Parameter: genopcoaxe

Explanation: Applies only if 'genopca' has been set as TRUE (✓ in *on server mode*). Indicate here the number of coordinates to be retained within the PCoA analysis. This number should always be lower than the number of genotypic variables or markers. This means that only the 'retained' principal coordinates will be shown in the results tables.

#### 9.4.1.40 Parameter: mantelt

Explanation: Indicate (with TRUE in *local mode* or ✓ in *on server mode*) if you wish to analyse the correlation matrix (Mantel, 1967) between the possible combinations of factors (ecogeographic vs. phenotypic vs. genotypic). All possible comparisons will be made according to whether phenotypic or genotypic data were entered or if an ecogeographic characterization matrix was created based on collection sites. A matrix of geographic distances will be generated for paired matrix comparisons.

#### 9.4.1.41 Parameter: mantelmeth

Explanation: Applies only if 'mantelt' has been set as TRUE (✓ in *on server mode*). Select the type of correlation to use in the Mantel test.

#### 9.4.1.42 Parameter: mantelper

Explanation: Applies only if 'mantelt' has been set as TRUE (✓ in *on server mode*). Enter the number of permutations you want to perform the Mantel test.

#### 9.4.1.43 Parameter: resultados (only for local mode)

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / instead of \ when indicating the path of the folder. For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 9.5. Results of DIVmapas

After defining all the parameters and paths (for *local mode*) that DIVmapas requires, click on the 'Run' (*local mode* in RStudio) or 'Start' (*on server mode*) buttons to start the analysis process of the tool.

After some time that may vary due to the type of analysis requested, DIVmapas will save the results in the path and folder specified in 'resultados' (in *local mode*), or in the User's Files and Results area in *on server mode*. Using DIVmapas can produce many results, which may be organized according to their data and/or their source analysis. Therefore, DIVmapas creates several folders within the path indicated in parameter 'resultados' (*local mode*) or in the User's Files and Results area (*on server mode*). The results are saved in the corresponding thematic folders that will be explained in the following sections. The point map corresponding to the collection sites will be saved outside these folders in two versions ('ShapefilePuntosPasaporte.shp' and 'mapa\_puntospas\_google.kml'). The Parametros.Parameters.DIVmapas.txt file, which contains a register of the parameters used, will also be saved elsewhere

### 9.5.1 'ClassicMultivariateResults\_pais' folder

This folder contains graphics (.wmf format) and tables (.txt and .xls) generated by multivariate analyses (cluster analysis and principal component/coordinate analysis). Depending on the data entered, the following terms will appear in the file names: 'ecogeographic' (from ecogeographic characterization), 'genotypic' (from genotypic characterization), and 'phenotypic' (from phenotypic characterization).

### **9.5.2 'EcogeographicResults pais' folder**

This folder will appear when an ecogeographic diversity map is requested (parameter 'ecogeo') The diversity map is offered in four different formats (as an image 'mapadiv\_ecogeo\_pais.png'; as a Google Earth map 'mapadiv\_ecogeo\_pais.kml'; DIVA-GIS 'mapadiv\_ecogeo\_pais.grd', and geotif 'mapadiv\_ecogeo\_pais.tif'). In these maps, the average ecogeographic distances from each cell's accession of the area of influence are shown in different colours. This is the Euclidean distance whose possible values range from 0 (when there is only one accession or all accessions were collected in identical environments) to infinity.

You will find the following tables:

**9.5.2.1 'tabla\_estadisticas\_mapadiv\_ecogeo.txt'**. This table shows the statistics for the ecogeographic diversity map in terms of distance, i.e., the mean, standard deviation, and the maximum and minimum distance values defined in the set of cells making up the map.

**9.5.2.2 'TablaVarEcogeograficaspais.txt'**. This table contains ecogeographic characterization data from the accessions analysed.

### **9.5.3 'PhenotypicResults pais' folder**

This folder appears when a phenotypic diversity map is requested (parameter 'phenotip') and the table with the corresponding data has been entered. This contains the diversity map in three different formats (as an image 'mapadiv\_phenot\_pais.png'; as a Google Earth map 'mapadiv\_phenot\_pais.kml', and DIVA-GIS 'mapadiv\_phenot\_pais.grd'). In these maps, the average phenotypic distances from each cell's accession of the area of influence are shown in different colours. The distance corresponds to 1- Gower's general similarity coefficient (1971) and has possible values from 0 (when there is only one single accession or all accessions have the same phenotype) up to 1 (maximum difference).

You will only find the 'tabla\_estadisticas\_mapadiv\_phenot' table with .txt and .xls. extensions. This table shows the statistics for the phenotypic diversity map in terms of distance, i.e., the mean, standard deviation, and the maximum and minimum distance values defined in the set of cells making up the map.

### **9.5.4 'GenotypicResults pais' folder**

This folder appears when a genotypic diversity map is requested (parameter 'genotip') and the table with the corresponding data has been entered.

Inside the folder you will find the following maps:

**9.5.4.1 'mapadiv\_GenotDistance\_pais'.** This corresponds to the map of genotypic diversity measured in average distances in four formats ('.png' image; Google Earth '.kml'; DIVA-GIS '.grd', and geotif '.tif'). In these maps, the average genotypic distances from the areas of influence of each cell are shown in different colours. The distance corresponds to the 1- Dice similarity coefficient (1945) and has potential values from 0 (when there is only one single accession or all accessions have the same phenotype) to 1 (maximum difference).

**9.5.4.2 'mapadiv\_GroupSize\_pais'.** This corresponds to the map for the number of accessions analysed by cell ('.png' image; Google Earth '.kml'; DIVA-GIS '.grd', and geotif '.tif'). In these maps, the number of accessions for the areas of influence of each cell is shown in different colours.

**9.5.4.3 'mapadiv\_NeisGeneDiversity\_pais'.** This corresponds to the map of genotypic diversity measured by Nei's diversity index (1987) in four formats ('.png' image; Google Earth '.kml'; DIVA-GIS '.grd', and geotif '.tif'). In these maps, the values of the diversity index obtained from the accessions characterized by the area of influence of each cell are shown in different colours.

**9.5.4.4 'mapadiv\_ProportionVariableMarkers\_pais'.** This corresponds to the map showing the proportion of polymorphic markers in four formats ('.png' image; Google Earth '.kml'; DIVA-GIS '.grd', and geotif '.tif'). In these maps, the proportion of polymorphic molecular markers obtained from the accessions characterized by the area of influence of each cell is shown in different colours.

You will find the following tables:

**9.5.4.5 'tabla\_estadisticas\_mapa\_GenotDistance.txt'.** This table shows the statistics for the ecogeographic diversity map in terms of the Dice distance (1945), i.e., the mean, standard deviation, and the maximum and minimum distance values defined in the set of cells making up the map.

**9.5.4.6 'tabla\_estadisticas\_mapa\_NeisGeneDiversity.txt'.** This table shows the statistics for the genotypic diversity map in terms of Nei's genetic diversity index (1987), i.e., the mean, standard deviation, and the maximum and minimum distance values for this index, defined in the set of cells making up the map.

### **9.5.5 'MantelCorrelationResults\_pais' folder**

All tables with the distance matrices calculated for all accessions simultaneously ('Matriz\_distancia\_') and those containing the results of Mantel's matrix correlation tests (1967) will be saved in this folder. The name of each table indicates the kind of comparison process made. Dice's distance matrix is used to measure correlations where genotypic data are involved. For example, the file 'Mantel\_genotypic\_Vs\_phenotypic.txt' contains the results of the correlation matrix between genotypic distances (Dice) and phenotypic distances (Gower). It is important to note that DIVmapas also calculates the matrix of geographical distances (calculated in decimal degrees) to enable matrices to be compared in terms of the geographical distance component.

## 9.6. References

Damme, P., Garcia, W., Tapia, C., Romero, J., Manuel Siguéñas, M., Hormaza, J.I. 2012. Mapping Genetic Diversity of Cherimoya (*Annona cherimola* Mill.): Application of Spatial Analysis for Conservation and Use of Plant Genetic Resources. PLoS ONE 7(1): e29845. doi:10.1371/journal.pone.0029845

Dice, L.R. 1945. Measures of the Amount of Ecologic Association Between Species. Ecology 26:297–302.

FAO, BIOVERSITY. 2015. FAO/Bioversity multi-crop Passport descriptors V.2. Disponible en <https://bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/>

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857:74.

Hoffmann, M.H., Glaß, A.S., Tomiuk, J., Schmuths, H., Fritsch, R.M., Bachmann, K. 2003. Analysis of molecular data of *Arabidopsis thaliana* (L.) Heynh. (Brassicaceae) with Geographical Information Systems (GIS). Molecular Ecology 12: 1007–1019

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209-220.

Thomas, E., van Zonneveld, M., Loo, J., Hodgkin, T., Galluzzi, G., van Etten, J. 2012. Present spatial diversity patterns of *Theobroma cacao* L. in the neotropics reflect genetic differentiation in pleistocene refugia followed by human-influenced dispersal. PLoS ONE 7(10): e47676. doi:10.1371/journal.pone.0047676

van Zonneveld M, Scheldeman X, Escribano P, Viruel MA, Van Damme P, et al. (2012) Mapping Genetic Diversity of Cherimoya (*Annona cherimola* Mill.): Application of Spatial Analysis for Conservation and Use of Plant Genetic Resources. PLoS ONE 7(1): e29845. doi:10.1371/journal.pone.0029845





**National Workshop, Esperanza, Santa Fe (Argentina), April 2016.**





# 10 | ColNucleo Tool

## 10.1. Concept of core collection

A core collection is a subset, or a fraction of an original collection, organized for any number of reasons. The size of the original collection is the key determining factor when deciding to create a core collection. Core collections are used as a solution when the size of the original collections becomes a problem. A larger collection is often a problem when multiplying, characterizing, or evaluating germplasm, particularly when economic resources are limited. The size also affects the selection of materials for breeding programs and the creation of active or working collections, for example. The definition of a 'large collection' depends on the conditions of each site and may range from 500 to 1,000, 2,000 or more accessions.

A core collection is usually made of 10% of the total accessions in the original collection, although there are studies that place the optimum percentage above or below this value (Parra Quijano *et al.*, 2011a). This percentage is known as a 'sampling intensity'.

The determination of a core collection should never jeopardize the conservation of non-selected accessions, known as the 'reserve collection'. A core collection can help to set priorities when resources are limited, and decisions need to be made about specific conservation activities; this does not exempt the user from their responsibility to conserve the collection in its entirety. For example, when you need to multiply germplasm using a core collection but with limited resources, you could begin by multiplying accessions from the core collection and perform another multiplication cycle for the rest of the collection with other additional resources.

Independently of the reasons for its creation, the main feature of a core collection, as compared with other kinds of subcollections, is that it should represent the genetic diversity contained in the original collection. This implies that a core collection should contain accessions that are as dissimilar as possible so that genetic duplicates or closely-related accessions are not included (Brown, 1995). Thus, to obtain a subset of genetically dissimilar accessions, it is essential to have information about the genetic composition of the collection, in other words, characterization data.

This is one of the first difficulties in obtaining core collections: when resources are limited, it may not be feasible to characterize a collection of over 1,000 or 2,000 accessions. Genotypic and phenotypic characterizations usually demand significant financial resources and human effort that many institutions are unable to afford or can only pay partially. However, other kinds of characterization data may be employed to overcome this problem.

In certain cases, when core collections were needed and no characterization data were available for this purpose, one solution proposed was to use passport data, in particular the administrative details describing the location of the collection site (country, state, province). The idea was to assimilate different administrative collection units into different environments to achieve a core collection representative of all administrative units and environments. Several administrative core collections were created in this way for species such as the peanut (Upadhyaya *et al.*, 2003), pigeon pea (Reddy *et al.*, 2005), sesame (Xiourong *et al.*, 2000), and sorghum (Grenier *et al.*, 2001). However, this kind of collection does not guarantee that the core collection includes the greatest variety of accessions in terms of the environment from which they were collected, as the different administrative units answer to man-made divisions and do not necessarily correspond to different environments.

### **10.1.1 Clustering strategy**

The first step in setting up a core collection is to organize the original collection into clusters according to affinity. As mentioned previously, a core collection requires ecogeographic, genotypic, and phenotypic data or, in the case of administrative core collections, passport data. This information is used to create clusters of similar or related accessions. Clusters can be created with multivariate classification methods using germplasm characterization data.

One option to create ecogeographic core collections proposes the use of ecogeographical land characterization where the germplasm occurs (ELC maps) instead of the usual germplasm characterization approach. Thus, accessions are grouped according to the ecogeographic category where they occur. This is helpful when new accessions are added to the core collection, as it becomes unnecessary to repeat the cluster analysis. All you need to know is to which cluster (ecogeographic map category) the new accession belongs (Parra Quijano *et al.*, 2011b).

### **10.1.2 Determination of quotas by allocation strategies**

Subsequently, the number of accessions to be selected for each affinity cluster is determined. This number or quota is determined by the allocation strategy selected by the curator as appropriate. As the use of core collections has become more widespread, an increasing number of allocation strategies have been proposed. The complexity and sophistication of these strategies have also increased over time. However, some comparative studies show that the most complex strategies do not necessarily produce the most representative core collections (Parra Quijano *et al.*, 2011b). The most popular, simple, and widely tested strategies are as follows (Yonezawa *et al.*, 1995):

1. Random (R): Accessions are randomly selected from the whole collection. Clusters created by stratification are ignored.
2. Constant (C): The same number of accessions is selected from each cluster, regardless of how many accessions it contains.
3. Proportional (P): The number of accessions selected from each cluster is proportional to its size (total number of accessions contained).
4. Logarithmic (L): The number of accessions selected from each cluster is proportional to the logarithm of its size (total number of accessions contained).
5. Diversity dependent (G): The number of accessions selected from each cluster is proportional to the diversity it represents. This strategy requires access to characterization data, in addition to the clusters generated by stratification.

### **10.1.3 Information about the availability of accessions**

Many scientific studies about the creation of core collections perform simulations to determine the best cluster and allocation strategy for producing the most representative core collection for each case, using the entire collection for this purpose. However, these theoretical approaches and simulations may produce core collections that in practice cannot be created as the selected accessions are unavailable. Several factors influence an accession's availability

for inclusion in a core collection, including the number of seeds available, if the accession is only represented in the base collection or if there are any restrictions conditioning its use and distribution. For this reason, it is important to consider the information about accession's availability the curator may provide when drawing up a core collection for practical purposes.

## 10.2. Ecogeographic core collections

Ecogeographic characterization is an alternative method of creating core collections. Considering the relationship between phenotype, genotype, and the environment, a core collection based on ecogeographic characterization may be representative in terms of the environmental conditions of the populations where the accessions originated. It may also be representative of their phenotypes and genotypes, provided that this representativeness is evaluated according to those phenotypic or genotypic traits related to adaptation (Parra Quijano *et al.*, 2011a).

The use of ecogeographic characterization data to establish core collections has been documented since 1995 when a core collection of *Phaseolus vulgaris* was created at the International Center for Tropical Agriculture (Centro Internacional de Agricultura Tropical - CIAT) (Tohme *et al.*, 1995). However, the wide availability of GIS could not be applied to plant genetic resources and ecogeographic information layers until the decade following the year 2000, and core ecogeographic collections did not reappear in the international scientific context until 2008, with the case of *Trifolium spumosum* (Ghamkhar *et al.*, 2008).

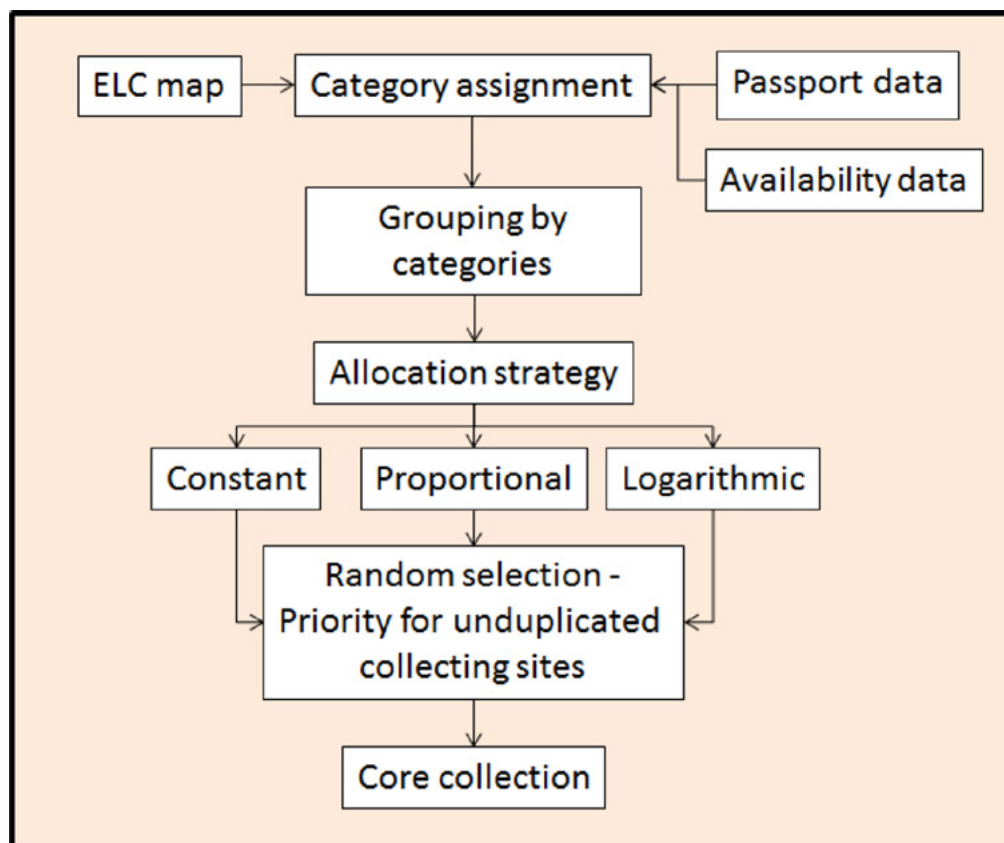
Subsequently, a couple of studies on different kinds of ecogeographic collections determined that the combination of an ELC map as a clustering strategy with a proportional map as an allocation strategy generated highly representative ecogeographic and phenotypic core collections for *Lupinus* spp. and *Phaseolus vulgaris*, respectively (Parra Quijano *et al.*, 2011a, 2011b). In these studies, up to 16 different combinations of clustering and allocation strategies generated similar or inferior results in terms of ecogeographic and phenotypic representativeness as compared with the combination of the ELC map with proportional allocation.

## 10.3. Obtaining ecogeographic core collections in ColNucleo

Following the recommendations of certain scientific studies on core collections and representativeness, ColNucleo tool enables ecogeographic core collections to be obtained using the combination of ELC map clusters with three allocation methods (C, P, and L). The ELC map should be generated using the 'ELCmapas' tool (see chapter on this tool).

Fig. 42 shows how the ELC map category corresponding to each accession's collection site, including coordinates, is extracted as a first step. The accessions are then grouped according to the ELC map category assigned. ColNucleo sets quotas or a number of accessions for each group making up the core collection according to the allocation strategy and sampling intensity selected by the user. ColNucleo then determines if the quota can be met by accessions without geographical duplicates (not necessarily genetic) designated as 'available' by the curator if the user has selected the option of using data about availability. Accessions without duplicates will have precedence over duplicate

accessions. If the quota is smaller than the number of non-duplicate accessions available, a random selection will be made from these accessions. If the quota is larger, all non-duplicate accessions will be selected, and the shortfall will be made up with a random selection of duplicate accessions. Finally, the selected accessions will be marked with the number 1 (one) in a new column added to the accessions' passport table. If only availability data is used, the core collections obtained may be incomplete if there are not enough accessions to represent one or more ELC categories. For this reason, ColNucleo generates an additional table showing the accessions that need to be made available for the core collection to represent all the ELC categories according to the quotas set.



**Figure 42.** Illustration of the process followed by ColNucleo tool to obtain ecogeographic core collections.

## 10.4. Format of passport table for ColNucleo

ColNucleo uses the FAO/Biodiversity 2012 passport table with modifications which, in turn, uses GEOQUAL, Representa, and ECOGEO tools with the addition of a field on the right side named 'AVAILAB' that determines the availability of each accession. Available accessions are coded with the number 1 (one) in column AVAILAB, unavailable accessions with a 0 (zero) and those for which there is no information are coded with the letters NA.

## 10.5. Using ColNucleo tool

Once the user has installed CAPFITOGEN3 *local mode* tools or accessed CAPFITOGEN3 *on server mode* and ColNucleo tool has been selected, a set of parameters must be defined. Note that for this tool to work correctly, the user must have obtained an ELC map using ELCmapas tool. The following files (produced by the ELCmapas tool) must be copied and pasted in the CAPFITOGEN3/ELCmapas path (*local mode*) or uploaded to the server when the respective parameters require it:

- mapa\_elc\_pais.grd
- mapa\_elc\_pais.gri
- estadist\_ELC\_pais.txt

Note that 'pais' is the name of the country or region for which the map was made.

### 10.5.1 Initial parameters defined by the user

#### 10.5.1.1 Parameter: ruta (only for local mode)

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### 10.5.1.2 Parameter: pasaporte

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area. This table is identical to the passport table that is used as a model for other CAPFITOGEN tools, but contains an additional column called 'AVAILAB'. This additional column indicates the availability of each accession to be selected for a core/nuclear collection.

#### 10.5.1.3 Parameter: geoqual

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte').

Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### 10.5.1.4 Parameter: **totalqual**

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

#### 10.5.1.5 Parameter: **mapaelc**

Explanation: Enter the name of the file (*local mode*) containing the ELC map generated by the ELCmapas tool or upload it to the server (*on server mode*). This map should be found in the CAPFITOGEN3/ELCmapas folder, one of the folders making up the CAPFITOGEN directory. The map should be in DIVAGIS format (.grd extension, exactly as generated by the ELCmapas tool) and its name should be entered with the file extension. Thus, if the name of the map is 'mapa\_elc\_spain', you should enter 'mapa\_elc\_spain.grd' (*local mode*) or search for this file (*on server mode*).

#### 10.5.1.6 Parameter: **statelec**

Explanation: Enter the name of the file (*local mode*) with the table of the ELC map's descriptive statistics generated by the ELCmapas tool or select it and upload it to the server (*on server mode*). The tool usually names this kind of file as 'Estadist\_ELC\_' plus the name of the country or region. Like the ELC map, this file should also be located in the CAPFITOGEN3/ELCmapas folder. Similarly, the name should be followed by the file extension (*local mode*), which in this case is '.txt' because the file is a table. Therefore, if the file is called 'Estadist\_ELC\_spain', the name should be 'Estadist\_ELC\_spain.txt'.

#### 10.5.1.7 Parameter: **distdup**

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites represent the same population. The value of zero (by default) excludes accessions with identical coordinates from the analysis of representativeness. The determination of the distance depends on biological (gene flow) and spatial (mean population sizes) conditions. This is a specific parameter for the target species, and it will often be necessary to consult an expert for his/her concept.

#### 10.5.1.8 Parameter: **porcol**

Explanation: This corresponds to the sampling intensity. Indicate the size required for the core collection expressed as a percentage of the size of the original collection (values from 0 to 100). For example, if the original collection contains 2,000 accessions and a core collection of 200 accessions is required, then enter '10'. For a core collection of 300 accessions, enter '15'.

### 10.5.1.9 Parameter: *estratcol*

Explanation: Select a strategy to set the allocation of representation quotas for each ecogeographic category of the ELC map. You may choose from these strategies: 'C' constant (using the same quota for all categories); 'P' proportional (quotas that are proportional to the number of accessions in each category), or 'L' logarithmic (quotas that are proportional to the logarithm of the number of accessions in each category).

### 10.5.1.10 Parameter: *avilab*

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to use the accession availability column to select accessions for a core collection. Remember that the passport table in this tool includes a column called 'AVAILAB' showing which accessions from the original collection are available to make up a core collection. Accessions may be marked 0 (unavailable), 1 (available), or NA (no information/unavailable). If you prefer not to use the information on availability, the selection of accessions will be carried out based on the total number of accessions. Availability is defined according to the curator's criteria and may be determined by the number of seeds preserved, their germination, or a range of other factors.

### 10.5.1.11 Parameter: *resultados* (only for *local mode*)

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 10.6. Results of ColNucleo

Once the analysis is finished, ColNucleo will produce a total of four tables in the path determined in parameter '*resultados*' (*local mode*) or in the User's Files and Results area (*on server mode*) if the user has indicated that germplasm availability data (column AVAILAB in the passport table) should be used. However, three tables will be generated if the user has decided not to use this information.

**10.6.1.1 'CoreCollection.txt/xls'.** This table contains the passport table with the accessions selected by ColNucleo for the ecogeographic core collection and an additional column on the right of the table called 'BGcat', indicating the group or category in the ELC map to which the accession belongs, according to its collection site.

**10.6.1.2 "'CoreCollect\_Properties.txt/xls'.** This table shows several parameters introduced by the user and that ColNucleo has used to establish the core collection. The fields included are: 'Allocation\_strategy'; 'Sample\_size' (sampling intensity percentage); 'Use\_availability\_data' (whether availability data is used or not); 'No\_access\_sampled' (the size of the core collection), and 'No\_access\_to\_be\_multiplied' (the number of unavailable accessions or those for which there are no availability data) (only when using availability data and when such data is needed to create a complete core collection).



**10.6.1.3 'CoreCollect\_stats.txt/xls'**. This table contains statistics for each ELC map category (identified in column 'ELC\_CAT'). It contains the following columns on the right of column ELC\_CAT: 'FREC\_W\_DUPL' indicates the number of accessions, including geographical duplicates, whose collection site falls within each category; 'FREC\_WO\_DUPL' indicates the same as the previous column without the geographical duplicates; 'Porcent\_W\_DUPL' indicates the percentage of accessions (including duplicates) in each category; 'FreqClass\_W\_DUPL' indicates the quartile classification of occurrence frequency in each category; 'Duplicates' indicates the number of duplicate geographic accessions per category; 'N\_Availab' indicates the number of total available accessions (duplicates and nonduplicates) per category; 'N\_AvailabWO' indicates the number of nonduplicate accessions available per category; 'Q\_Even' or 'Q\_Prop' or 'Q\_Log' (the column heading depends on the allocation method selected) refer to the quota (the number of accessions that should represent each category *a priori*), and lastly 'CCfinal' indicates the number of accessions making up the ecogeographic core collection obtained by the ColNucleo tool based on the parameters entered and (when applicable) the availability of accessions.

**10.6.1.4 'EntriesToBeMultiplied.txt/xls'**. This table has the same column structure as 'CoreCollection.txt/xls' except that 'EntriesToBeMultiplied.txt/xls' shows the accessions selected by ColNucleo as part of the core collection but that are unavailable. Given the name of the file, it is assumed that these accessions need to be multiplied to become available for the core collection. However, there may be several reasons why they are unavailable, as explained in paragraph 10.1.3.

## 10.7. References

- Brown, A.H.D. 1995. The core collection at the crossroads. p. 3–19. In: Hodgkin, T., Brown, A.H.D., Hintum, T.J.L., Morales, E.A.V. (eds.) Core collections of plant genetic resources. John Wiley & Sons, New York, NY.
- Ghamkhar, K., R. Snowball, B.J. Wintle, Brown, A.H.D. 2008. Strategies for developing a core collection of bladder clover (*Trifolium spumosum* L.) using ecological and agro-morphological data. Australian Journal of Agricultural Research 59:1103–1112.
- Grenier, C., Hamon, P., Bramel-Cox, P.J. 2001. Core collection of sorghum: II. Comparison of three random sampling strategies. Crop Science 41: 241-246.
- Parra Quijano, M., Iriondo, J.M., Torres, M.E., De la Rosa, L. 2011a. Evaluation and validation of ecogeographical core collections using phenotypic data. Crop Science 51: 694-703.
- Parra-Quijano, M., Iriondo, J.M., de la Cruz, M., Torres, M.E. 2011b. Strategies for the development of core collections based on ecogeographical data. Crop Science 51: 656-666.
- Reddy, L.J., H.D. Upadhyaya, C.L.L. Gowda, S. Singh. 2005. Development of core collection in pigeonpea (*Cajanus cajan* (L.) Millspaugh) using geographic and qualitative morphological descriptors. Genetic Resources and Crop Evolution 52: 1049-1056.

Tohme, J., Jones, P., Beebe, S., Iwanaga, M. 1995. The combined use of agroecological and characterisation data to establish the CIAT *Phaseolus vulgaris* core collection. p. 95-107. In Hodgkin, T., Brown, A.H.D., Hintum, T.J.L., Morales, E.A.V. (eds.) Core collections of plant genetic resources. John Wiley & Sons, New York, NY.

Upadhyaya, H.D., Ortiz, R., Bramel, P.J., S. Singh, S. 2003. Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. *Genetic Resources and Crop Evolution* 50: 139-148.

Xiurong, Z., Yingzhong, Z., Yong, C., Xiangyun, F., Qingyuan, G., Mingde, Z., Hodgkin, T. 2000. Establishment of sesame germplasm core collection in China. *Genetic Resources and Crop Evolution* 47: 273-279.

Yonezawa, K., Nomura, T., Morishima, H. 1995. Sampling strategies for use in stratified germplasm collections. p. 35-53. In: Hodgkin, T., Brown, A.H.D., Hintum, T.J.L., Morales, E.A.V. (eds.) Core collections of plant genetic resources. John Wiley & Sons, New York, NY.





**Regional Workshop, Pretoria (South Africa), April 2015.**



# 11 | FIGS\_R Tool

## 11.1. Focused Identification of Germplasm Strategy

The technique used to select germplasm for practical purposes known as a 'Focused Identification of Germplasm Strategy' or FIGS, comes from a concept originally developed by Mackay (1990).

It seeks to identify accessions in a collection that could potentially be used by breeders. The potential for use in breeding is based on the ecogeographic information of collection sites and its association with traits of interest for breeders (Mackay and Street, 2004).

As FIGS uses abiotic ecogeographical variables to select germplasm, the association between ecogeographic variables and traits of interest for breeding is direct if the trait of interest is abiotic, or indirect if the trait is biotic. So, if a breeder is looking for germplasm with breeding potential and the trait of interest is the adaptation to drought conditions, he/she will directly look for germplasm from a collection location with low rainfall. If the trait of interest is biotic, such as the resistance to a pathogen, a relationship between a series of ecogeographic variables and the resistance to the pathogen needs to be established first. This will enable the subsequent selection of germplasm from a collection site whose ecogeographic conditions are associated with resistance to the pathogen.

There are two techniques for selecting germplasm using FIGS. The first is filtering accessions and the second is a calibration technique.

The filtering technique selects accessions from an ecogeographically characterized collection and chooses those that comply with certain values or ranges for the variables characterized. Sometimes what is selected is just a fraction of the distribution of an ecogeographic variable in the collection characterized. The values and ranges or the fraction of distribution, as well as the ecogeographic selection variable, are set by the researcher, curator, or breeder based on their knowledge of the species, the ecogeographic variable, and the trait of interest. An example of the application of this method was the indirect selection made by FIGS for a wheat strain resistant to the pest *Eurygaster integriceps* (El Bouhssini *et al.*, 2009). Another case is the direct application of FIGS used to identify genetic resources of *Vicia faba* able to adapt to drought conditions (Khazaei *et al.*, 2013). The calibration technique requires the entire collection (or almost all of it) to have been ecogeographically characterized (using accessions with coordinates). Additionally, it must have been evaluated at least partially for the trait of interest. The calibration technique takes place in two phases. In the first phase, mathematical and statistical analyses are used to establish the relationship between the presence or absence of the trait of interest and one or more ecogeographical variables. Once this relationship has been established, the presence or absence of the trait of interest is predicted from the non-evaluated fraction of the collection, using ecogeographic information available for the entire collection for this purpose. The prediction indicates which accessions would be potentially relevant to crop breeding. The application of the calibration technique can be seen in the studies of Endresen and his team for barley and wheat (Endresen, 2010; Endresen *et al.*, 2012).

The calibration technique lends itself naturally to indirect FIGS while the filtering technique can be used for both types. The calibration technique is methodologically more complex than the filtering technique, and its results are also assumed to be more accurate for detecting accessions with the trait of interest. However, the calibration tech-

nique has a drawback as it depends on partial collection evaluation data, which must also be sufficiently reliable to enable a valid relationship to be established between the ecogeographic variable and the trait of interest. This means that its application is restricted to 22% of the collections, which is the percentage of national collections including some form of biotic evaluation from 40 countries, according to the Second Report on the Status of Plant Genetic Resources for Food and Agriculture (FAO, 2010).

Regardless of the way that the FIGS subset is obtained, it should be validated with adaptation, tolerance, or resistance tests to ensure that the accessions selected do possess the trait of interest they were chosen for using the ecogeographical conditions of their collection sites.

## 11.2. FIGS subsets and core collections

A FIGS subset is the set of accessions with potential for use in breeding a cultivated species and that comes from a FIGS selection process.

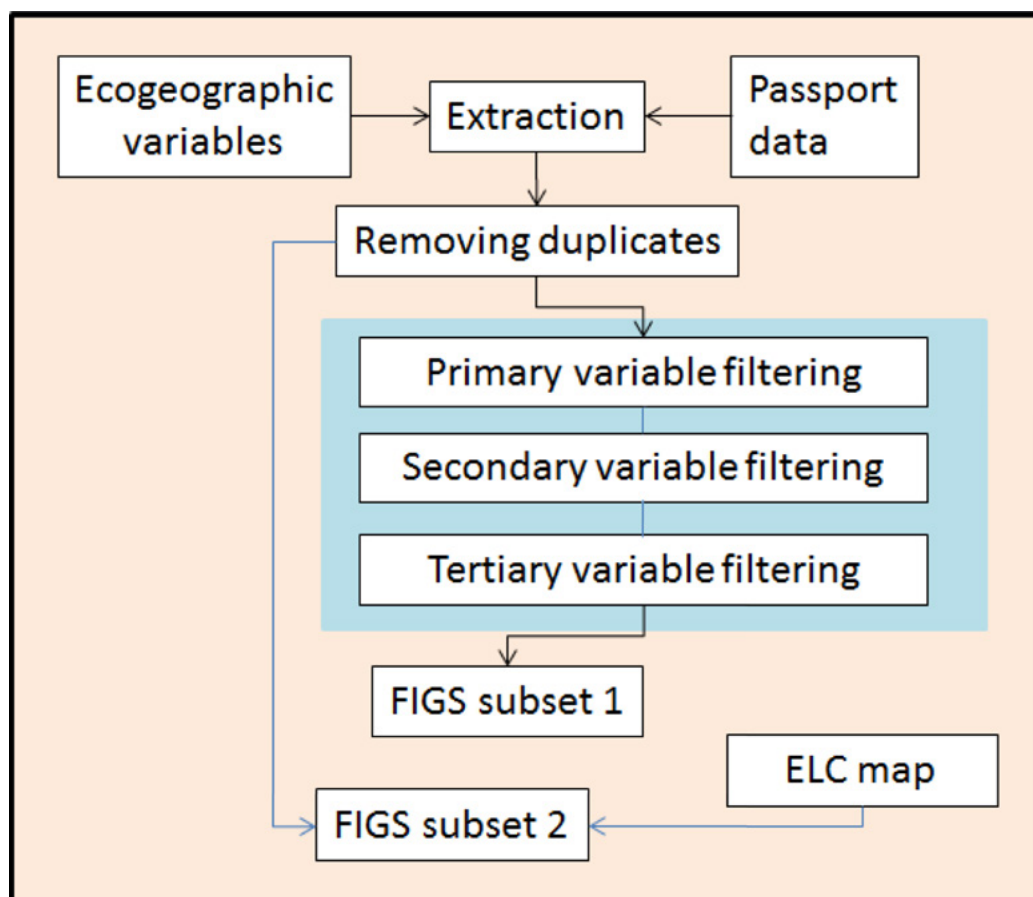
FIGS subsets, unlike a core collection, do not necessarily need to be representative of the variability of the original collection. A conventional FIGS subset carries a pronounced bias when selected: the interest of crop breeders. Thus, it is unlikely to be highly representative.

Another difference between a core collection and a FIGS subset is that as many of the latter may be established for a given species as there are traits of interest. In contrast, only one core collection is usually established per species.

However, as with core collections, establishing one or more FIGS subsets should not jeopardize the conservation of non-selected accessions. For example, while a core collection is used to prioritize the characterization and evaluation of specific accessions in a collection when there are no resources to do this for the entire collection, a FIGS subset seeks to enhance the use of a germplasm collection, by helping crop breeders (who are its main users) to locate material with the potential for integration into breeding programs.

## 11.3. Obtaining FIGS subsets using FIGS\_R tool

FIGS\_R tool can be used to obtain a FIGS subset using the filtering technique. FIGS\_R allows up to three selection variables to be used in hierarchical order. A primary variable (required) is used for the first filtering process, the secondary one (optional) filters the subset resulting from the first filter, and a tertiary variable (optional, and only used after the secondary variable) that filters the subset generated by the second filter. Any one of the 177 ecogeographical variables can be chosen as the primary, secondary, or tertiary variable, which are available in CAPFITOGEN3 tools. Fig. 43 shows the process followed by FIGS\_R to create FIGS subsets.



**Figure 43.** Illustration of the process followed by FIGS\_R tool to obtain FIGS subsets.

When using FIGS\_R, you can set the criteria for each selection variable. The first determines the range of values the accession must meet to be included in the FIGS subset. The second determines a specific fraction of the collection whose accessions have higher or lower values than the selection variable.

FIGS\_R employs some of the terms or definitions used in crop breeding, such as selection intensity and selection differential. Selection intensity defines the percentage of the initial collection to be included in the FIGS subset. Selection differential refers to the difference between the mean of the original collection and the mean of the FIGS subset for the selection variable(s).

Additionally, FIGS\_R tool can be used to create FIGS subsets that are ecogeographically balanced. In other words, if an ELC map has been created (with ELCmapas tool) using the second selection criteria (fraction of the collection), one may do the following: 1. Assign categories to each accession based on the ELC map category of the site collection, and 2. select the fraction of accessions with the highest or lowest values to define the selection variable for each category. Creating this kind of balance with an ELC map generally results in FIGS subsets with greater ecogeographic representativeness that are still useful for breeding programs, given their trait of interest.



Finally, please note that FIGS\_R tool can also work with information on the availability of accessions for selection. It also uses the same data accession format (passport data) as ColNucleo, i.e., the GEOQUAL format with the addition of the field 'AVAILAB'.

## 11.4 Using FIGS\_R Tool

Once the user has installed CAPFITOGEN3 *local mode* tools or accessed CAPFITOGEN3 *on server mode* and FIGS\_R tool has been selected, a set of parameters must be defined. Note that for this tool to work correctly, the user must have selected parameter 'mapaelc' and obtained an ELC map using ELCmapas tool. The following files (produced by the ELCmapas tool) must be copied and pasted in the CAPFITOGEN3/ELCmapas path:

- mapa\_elc\_pais.grd
- mapa\_elc\_pais.gri
- Estadist\_ELC\_pais.txt

Note that 'pais' is the name of the country or region for which the map was made. If the user does not select parameter 'mapaelc', he/she does not require an ELC map and can easily filter the germplasm with FIGS\_R.

### 11.4.1 Initial parameters defined by the user

#### 11.4.1.1 Parameter: ruta (only for local mode)

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### 11.4.1.2 Parameter: pais

Explanation: Select the country/region where all or most of the accessions you wish to analyse were collected. If accessions have been collected from more than one country, you may select a region, subcontinent, or continent. You can also use rLayer tool to produce your own work frames and select them here.

#### 11.4.1.3 Parameter: pasaporte

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only

have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area. This table is identical to the passport table, which is used as a model for other CAPFITOGEN tools, but contains an additional column called 'AVAILAB'. This additional column indicates the availability of each accession to be selected for a FIGS subset.

#### 11.4.1.4 Parameter: geoqual

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### 11.4.1.5 Parameter: totalqual

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

#### 11.4.1.6 Parameter: controlelc

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to use a previously created ELC map to determine the distribution of accessions in the FIGS subset concerning the map's different categories. For advanced users, this option enables them to obtain an additional FIGS subset in which accessions are selected for each ELC category. This requires the use of methods that make a selection from distribution fractions for all variables considered.

#### 11.4.1.7 Parameter: mapaclc

Explanation: Applies only if 'controlelc' has been set as TRUE (✓ in *on server mode*). Enter the name of the file (*local mode*) containing the ELC map generated by the ELCmapas tool or upload it to the server (*on server mode*). This map should be found in the CAPFITOGEN3/ELCmapas folder (*local mode*). The map should be in DIVAGIS format (.grd extension, exactly as generated by the ELCmapas tool) and its name should be entered with the file extension. Thus, if the name of the map is 'mapa\_elc\_spain', you should enter 'mapa\_elc\_spain.grd' (*local mode*) or search for this file (*on server mode*).

#### 11.4.1.8 Parameter: statelc

Explanation: Applies only if 'controlelc' has been set as TRUE (✓ in *on server mode*). Enter the name of the file (*local mode*) with the table of the ELC map's descriptive statistics generated by the ELCmapas tool or select it and upload it

to the server (*on server mode*). The tool usually names this kind of file as 'Estadist\_ELC\_' plus the name of the country or region. Like the ELC map, this file should also be located in the CAPFITOGEN3/ELCmapas folder. Similarly, the name should be followed by the file extension (*local mode*), which in this case is '.txt' because the file is a table. Therefore, if the file is called 'Estadist\_ELC\_spain', the name should be 'Estadist\_ELC\_spain.txt'.

#### 11.4.1.9 Parameter: distdup

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites represent the same population. The value of zero (by default) excludes accessions with identical coordinates from the analysis of representativeness. The determination of the distance depends on biological (gene flow) and spatial (mean population sizes) conditions. This is a specific parameter for the target species, and it will often be necessary to consult an expert for his/her concept.

#### 11.4.1.10 Parameter: availab

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to use the accession availability column to select accessions for a FIGS subset. This implies giving priority to available accessions but not restricting the possible consideration of those that are unavailable. Remember that the passport table in this tool includes a column called 'AVAILAB' showing which accessions from the original collection are available to make up a core collection. Accessions may be marked 0 (unavailable), 1 (available), or NA (no information/unavailable). Availability is defined according to the curator's criteria and may be determined by the number of seeds preserved, their germination or a range of other factors.

#### 11.4.1.11 Parameter: soloavailab

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to restrict the selection of accessions destined for the FIGS subset exclusively to the accessions designated as available (value 1 in the field 'AVAILAB').

#### 11.4.1.12 Parameter: resol1

Explanation: Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution but requires greater computing capacity than 5x5 km; however, this is not as limiting a factor as it is for ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents, or continents.

#### 11.4.1.13 Parameter: buffy

Explanation: Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish ecogeographic information to be extracted from a circular area around the collection site. Leaving this option unchecked means that information is extracted only from the point indicated by the collection site coordinates. This option is very useful, for example,

when most of the collection sites correspond to markets and not directly to the growers' farms, or when the actual location of the collection sites is uncertain even though GEOQUAL can grant high georeferencing quality. This feature is explained in the chapter dedicated to ECOGEO tool.

#### 11.4.1.14 Parameter: **tamp**

Explanation: Applies only if 'buffy' has been set as TRUE (✓ in *on server mode*). Specify the radius (in km) of a circular area around the point indicated by the collection site coordinates from which the ecogeographical information is to be extracted. The values extracted from the circular area will be averaged to obtain a single value. The information will be extracted from those cells whose centroid is within the circular area.

#### 11.4.1.15 Parameter: **variab1v**

Explanation: Select one (1) primary ecogeographical variable for which you wish to select accessions to obtain a FIGS subset. If you choose to select accessions based on one or two additional variables (secondary and tertiary variables), the variable selected at this point will be used for the first filter.

#### 11.4.1.16 Parameter: **variab1rang**

Explanation: Applies only if 'variab1cola' has been set as FALSE (□ in *on server mode*). Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to select accessions for the primary variable using a range of values, i.e., indicating minimum and maximum values to determine the range that will be used to select accessions for the FIGS subset.

#### 11.4.1.17 Parameter: **variab1min**

Explanation: Applies only if 'variab1rang' has been set as TRUE (✓ in *on server mode*). Specify the minimum value for the primary variable to determine the range to be used to select accessions for the FIGS subset.

#### 11.4.1.18 Parameter: **variab1max**

Explanation: Applies only if 'variab1rang' has been set as TRUE (✓ in *on server mode*). Specify the maximum value for the primary variable to determine the range required to select accessions for the FIGS subset.

#### 11.4.1.19 Parameter: **variab1cola**

Explanation: Applies only if 'variab1rang' has been set as FALSE (□ in *on server mode*). Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to select accessions for the primary variable using a distribution fraction, i.e., a percentage of the original collection whose values are either higher or lower than the primary variable.

#### 11.4.1.20 Parameter: **variab1vpor**

Explanation: Applies only if 'variab1cola' has been set as TRUE (✓ in *on server mode*). Determine the distribution fraction (as a percentage) that you wish to select to make up the FIGS subset. The values allowed range from 0 to 100.

#### 11.4.1.21 Parameter: **variab1vhl**

Explanation: Applies only if 'variab1cola' has been set as TRUE (✓ in *on server mode*). Select the distribution fraction you wish to select for the primary variable.

#### 11.4.1.22 Parameter: **variab2**

Explanation: Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to use a secondary variable to select accessions for a FIGS subset. The values of this variable will be used to select the accessions from the subset that was previously selected using the primary variable.

#### 11.4.1.23 Parameter: **variab2v**

Explanation: Applies only if 'variab2' has been set as TRUE (✓ in *on server mode*). Select one (1) secondary ecogeographic variable you wish to use to select accessions for a FIGS subset. It may be the same as the primary variable.

#### 11.4.1.24 Parameter: **variab2rang**

Explanation: Applies only if 'variab2' has been set as TRUE (✓ in *on server mode*) and variab2cola has been set as FALSE (□ in *on server mode*). Mark this option this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to select accessions for the secondary variable using a range of values, i.e., by indicating minimum and maximum values to determine the range that will be used to select accessions for the FIGS subset.

#### 11.4.1.25 Parameter: **variab2min**

Explanation: Applies only if 'variab2rang' has been set as TRUE (✓ in *on server mode*). Specify the minimum value for the secondary variable to determine the range to be used to select accessions for the FIGS subset.

#### 11.4.1.26 Parameter: **variab2max**

Explanation: Applies only if 'variab2rang' has been set as TRUE (✓ in *on server mode*). Specify the maximum value for the secondary variable to determine the range for selecting accessions for the FIGS subset.

**11.4.1.27 Parameter: variab2cola**

Explanation: Applies only if 'variab2' has been set as TRUE (✓ in *on server mode*) and variab2rang has been set as FALSE (□ in *on server mode*). Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to select accessions for the secondary variable using a fraction of the remaining distribution, i.e., a percentage of the subset selected by the primary variable with the highest or lowest values compared to the secondary variable.

**11.4.1.28 Parameter: variab2vpor**

Explanation: Applies only if 'variab2cola' has been set as TRUE (✓ in *on server mode*). Determine the fraction of the remaining distribution (as a percentage) that you wish to select to make up the FIGS subset using the secondary variable. The values allowed range from 0 to 100.

**11.4.1.29 Parameter: variab2vhl**

Explanation: Applies only if 'variab2cola' has been set as TRUE (✓ in *on server mode*). Select the distribution fraction you wish to select for the secondary variable.

**11.4.1.30 Parameter: variab3**

Explanation: Applies only if 'variab2' has been set as TRUE (✓ in *on server mode*). Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to use a tertiary variable to select accessions for a FIGS subset. The values of this variable will be used to select accessions from the subset previously selected using the primary and secondary variables. If the use of a secondary variable has not been previously determined, the selection of a tertiary variable will not affect the composition of a FIGS subset.

**11.4.1.31 Parameter: variab3v**

Explanation: Applies only if 'variab3' has been set as TRUE (✓ in *on server mode*). Select one (1) tertiary ecogeographic variable you wish to use to select accessions for a FIGS subset. This may be the same as the primary or secondary variable.

**11.4.1.32 Parameter: variab3rang**

Explanation: Applies only if 'variab3' has been set as TRUE (✓ in *on server mode*) and variab3cola has been set as FALSE (□ in *on server mode*). Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to select accessions for the tertiary variable using a range of values, i.e., by indicating minimum and maximum values to determine the range that will be used to select accessions for the FIGS subset.

#### 11.4.1.33 Parameter: **variab3min**

Explanation: Applies only if 'variab3rang' has been set as TRUE (✓ in *on server mode*). Specify the minimum value for the range of the tertiary variable to be used to select accessions for the FIGS subset.

#### 11.4.1.34 Parameter: **variab3max**

Explanation: Applies only if 'variab3rang' has been set as TRUE (✓ in *on server mode*). Specify the maximum value for the range of the tertiary value to be used to select accessions for the FIGS subset.

#### 11.4.1.35 Parameter: **variab3cola**

Explanation: Applies only if 'variab3' has been set as TRUE (✓ in *on server mode*) and variab3rang has been set as FALSE (□ in *on server mode*). Mark this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to select accessions for the tertiary variable using a fraction of the remaining distribution, i.e., a percentage of the subset selected by the primary and secondary variables whose values are higher or lower than the tertiary variable.

#### 11.4.1.36 Parameter: **variab3vpor**

Explanation: Applies only if 'variab3cola' has been set as TRUE (✓ in *on server mode*). Determine the fraction of the remaining distribution (as a percentage) that you wish to select for the FIGS subset using the tertiary variable. The values allowed range from 0 to 100.

#### 11.4.1.37 Parameter: **variab3vhl**

Explanation: Applies only if 'variab3cola' has been set as TRUE (✓ in *on server mode*). Select the distribution fraction you wish to select for the tertiary variable.

#### 11.4.1.38 Parameter: **resultados (only for local mode)**

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 11.5. Results of FIGS\_R

After defining all the parameters and paths (for *local mode*) that FIGS\_R requires, the analysis process of the tool will start by clicking on the 'Run' button (*local mode* in RStudio) or 'Start' (*on server mode*). FIGS\_R will produce between three and five tables.

When an ELC map is not included in the analysis (parameter 'controlelc'), only the following three tables appear:

### **11.5.1 'FIGS regular.txt/xls'**

This table contains the identification of the accessions selected for the FIGS subset (field 'ACCENUMB') as well as the site collection coordinates ('DECLATITUDE' and 'DECLONGITUDE'), the field of availability ('AVAILAB') and includes as many columns as the number of selection variables used.

### **11.5.2 'FIGS stat table.txt/xls'**

This table summarizes the characteristics of both the original collection and the FIGS subset. It uses statistics on the intensity of the selection achieved, as well as the selection average, and the maximum, minimum, and differential selection values for each selection variable.

### **11.5.3 'Passport FIGS R.txt/xls'**

This is the passport table introduced by the user into the analysis. It has an additional field for each selection variable called 'SEL\_VAR' followed by the numbers 1, 2, or 3. In these fields, accessions included in the FIGS subset are marked '1' while those not selected are marked 'NA'. Thus, the selection process using the primary variable (defined in parameter 11.4.1.15) selects the accessions identified with a '1' in the field 'SEL\_VAR1'. If a secondary variable is used, the accessions selected during the second filtering process are identified with a '1' in the field 'SEL\_VAR2'. Finally, if a tertiary variable is used, the accessions selected during the third filtering process are identified with a '1' in the field 'SEL\_VAR3'.

When an ELC map is included to provide more information about the ecogeographical characteristics of a FIGS subset, a new table appears:

### **11.5.4 'FIGS freq ELCmap.txt/xls'**

This table shows frequency values as well as the number of duplicates and accessions available for each ecogeographic category like the ColNucleo table ('CoreCollect\_stats.txt/xls'). On the left side of the table there are also three new fields identified with the prefix 'FIGS\_var' and then the numbers 1, 2, or 3. Thus, the number of accessions selected by the primary variable for each ELC category appears in the field 'FIGS\_var1'; the number of accessions selected by the secondary variable in the second filtering process performed for each ELC category appears in 'FIGS\_var2', and the number of accessions selected by the tertiary variable in the third filtering process for each ELC category appears in 'FIGS\_var3'.

Finally, if only the second selection method (collection fraction) has been used for the primary, secondary, and tertiary selection variables – meaning that the options 'variab1cola', 'variab2cola' and 'variab3cola' have been checked (parameters 11.4.1.19, 11.4.1.27, and 11.4.1.35, respectively) – then the results will include a fifth table:



### **11.5.5 'FIGS UnderELC.txt/xls'**

This table includes the same fields as in 'Passport\_FIGS\_R.txt' (section 11.5.3), but in this case, it contains only those accessions from the FIGS subset balanced by the ELC map. These accessions also include the fields 'SEL\_VAR1', 'SEL\_VAR2', and 'SEL\_VAR3' marked with a '1' to indicate whether these accessions would have also been selected for a FIGS without using an ELC map. On the left side of the table, there will be up to three new fields called 'var\_eco1', 'var\_eco2' and 'var\_eco3', depending on how many selection variables have been used. These fields will show the values for the selection variables extracted from each collection site ('var\_eco1' for the primary variable values, 'var\_eco2' for the secondary variable values, and 'var\_eco3' for the tertiary variable values).

Additionally, the table 'FIGS\_freq\_ELCmap.txt/xls' (section 11.5.4) will include up to three new fields on the left side, under the headings 'No\_by\_var1', 'No\_by\_var2' and 'No\_by\_var3'. These fields show the number of accessions selected for the FIGS subset balanced by the ELC map in each selection process: 'No\_by\_var1' for the first filtering process using the primary variable, 'No\_by\_var2' for the second filtering process using the secondary variable, and 'No\_by\_var3' for the third filtering process using the tertiary variable.

## **11.6. References**

El Bouhssini, M. E., Street, K., Joubi, A., Ibrahim, Z., Rihawi, F. 2009. Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. *Genetic Resources and Crop Evolution* 56: 1065– 1069.

Endresen, D.T.F. 2010. Predictive association between trait data and ecogeographic data for Nordic barley landraces. *Crop Science* 50: 2418-2430.

Endresen, D.T.F., Street, K., Mackay, M., Bari, A., Amri, A., De Pauw, E., Nazari, K., Yahyaoui, A. 2012. Sources of resistance to stem rust (Ug99) in bread wheat and durum wheat identified using Focused Identification of Germplasm Strategy. *Crop Science* 52: 764-773.

FAO 2010. *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome

Khazaei, H., Street, K., Bari, A., Mackay, M., Stoddard, F.L. 2013. The FIGS (Focused Identification of Germplasm Strategy) approach identifies traits related to drought adaptation in *Vicia faba* genetic resources. *PLoS One* 8(5): e63107. doi: 10.1371/journal.pone.0063107

Mackay, M.C. 1990. Strategic planning for effective evaluation of plant germplasm. p. 21-25 In: Srivastava, J.P., Damania, A.B. (eds.). *Wheat genetic resources: Meeting diverse needs*. John Wiley & Sons, Chichester, UK.

Mackay, M. C., Street, K. 2004. Focused identification of germplasm strategy – FIGS. p. 138-141. In: Black, C.K., Panozzo, J.F., Rebetzke, G.J. (eds.). *Cereals 2004. Proceedings of the 54th Australian Cereal Chemistry Conference*

and the 11th Wheat Breeders' Assembly, 21-24 September 2004, Canberra, Australian Capital Territory (ACT).  
Cereal Chemistry Division, Royal Australian Chemical Institute, Melbourne, Australia.





**National Workshop, Colonia (Uruguay), April 2014.**



# 12 | rLayer Tool

## 12.1. Creating ‘tailored’ information layers

Frequently, the distribution range of the germplasm collection or the population occurrence data of a species are beyond the boundaries of a country. Sometimes such distribution ranges are very small and do not even reach the boundaries of a province (some of the smallest administrative divisions). These scenarios can cause problems when trying to use CAPFITOGEN tools since the ecogeographic information (layers) is organized by countries, although there are some exceptions in which global, subcontinent (Central or South America), or continent (Europe) coverage is provided.

The rLayer tool has been specifically designed for such problematic cases. The user will be able to design sets of information layers (for up to 177 variables, excluding latitude and longitude which are not necessary for the CAPFITOGEN analysis of GIS layers) covering the precise distribution of their collections or individuals/populations of their passport data, regardless of whether they exceed the boundaries of the country or not.

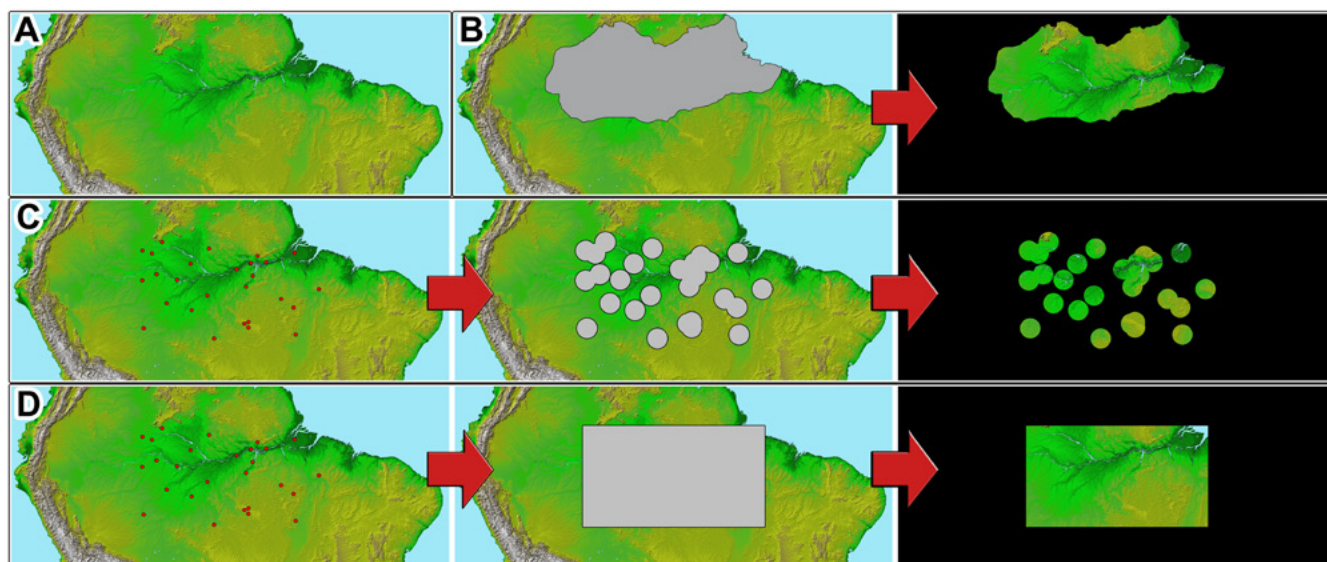
These sets are very useful to carry out more realistic, adjusted, or precise analyses with tools such as ELCmapas, ECOGEO, DIVmapas, ColNucleo, FIGS\_R, or Modela.

## 12.2. How does rLayer work?

From CAPFITOGEN3, rLayer tool creates new sets of ecogeographic information layers by cropping the global coverage layers to the actual size of the distribution range of collections or the user’s population occurrence data. This process is carried out in three ways that are described below and are illustrated in Fig. 44:

- ‘polygon’ method: The user provides the tool with a vector map or shapefile, usually composed of a single polygon that will be used as a template to crop the 177 layers of ecogeographic information. This template map must be in the WGS84 coordinate system and must represent land areas greater than the resolution selected for the set of remembered ecogeographic layers (1x1, 5x5, 10x10, or 20x20 km).
- ‘square’ method: The template to cut the layers of ecogeographic information results from the distribution range of collection sites or the occurrence of populations from the passport table. In this case, a rectangle is traced, and each side coincides with the maximum and minimum latitude and longitude values of the distribution.
- ‘buffer’ method: The template for cropping the layers of ecogeographic information results from circular areas around each collection site or population occurrence from the passport table. This area is plotted from a radius value that the user defines. Overlapping circular areas create contiguous areas.

The user must indicate in the tool the resolution wanted for the new set of layers, which should match the available downloaded resolutions (1x1, 5x5, 10x10, or 20x20 km). The user should also indicate the name of the new set of layers choosing amongst ‘user1’, ‘user2’ or ‘user3’ (*local mode*) or any name different from that of a country (*on server mode*). A very important feature of rLayer in its update for CAPFITOGEN3 is that the layers it crops and produces are in .tif format instead of .grd format. This results in a saving in disk space.



**Figure 44.** Illustration of a layer cropping process performed by rLayer. In each case, the result of this process is observed on a dark background. A. Original global layer (1x1 km) for elevation content (altitude) to be cropped. B. Cropping process using a shapefile containing a polygon supplied by the user ('polygon' method). C. Cropping process based on circular areas traced around germplasm collection sites or observed populations ('buffer' method). D. Cropping process based on a rectangle. The sides are traced according to the maximum and minimum values of latitude and longitude of the points of the distribution range of collection sites or observed populations ('square' method).

Once the process has finished, the user can select the new set of layers under parameter 'pais' in tools such as ECO-GEO, DIVmapas, ColNucleo, or FIGS\_R. Regardless of the name chosen by the user, under parameter 'pais' the three options given by rLayer (i.e., 'user1', 'user2', and 'user3') will be shown at the end of the list of countries and regions in *local mode*. Either in *local mode* or *on server mode*, the user must remember the name chosen in rLayer and then select it, if he/she wants to use the new set of layers.

Similarly, the user must remember the available resolutions of the downloaded and unzipped files as the size of the cell will be required in parameter 'resol1'. This parameter will appear in the tools that use ecogeographic information layers.

For the 'square' method, rLayer can identify both, the maximum and minimum latitude and longitude values of the collection's distribution or occurrence data, previously inserted by the user into parameter 'pasaporte'. According to the spatial resolution of the global layers, the following frames are created: 1x1 km = 0.05 decimal degrees (c.a. 6 km on the Equator), 5x5 km = 0.10 decimal degrees (c.a. 11 km on the Equator), 10x10 km = 0.15 decimal degrees (c.a. 17 km on the Equator) and 20x20 km = 0.20 decimal degrees (c.a. 22 km on the Equator). With the final minimum and maximum values (plus or minus the frames), rLayer crops each of the 177 global ecogeographic information layers stored in the 'world' folder (path 'CAPFITOGEN2/rdatamaps/world'). The resulting cropped layers are saved under the chosen name ('user1', 'user2' or 'user3') in a folder within the same path ('CAPFITOGEN2/rdatamaps/

world') for *local mode*. For *on server mode*, the variables will be available to the user in the parameter 'pais', from the drop-down list with country names. There, the user must find the name that he/she assigned to the set of custom cropped ecogeographic layers when executing rLayer tool.

## 12.3. Using rLayer tool

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and rLayer tool has been selected, the user should specify a series of parameters.

### 12.3.1 Initial parameters defined by the user

#### 12.3.1.1 Parameter: ruta (only for local mode)

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example: F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### 12.3.1.2 Parameter: cropway

Explanation: Select one of three methods for cropping layers: 'polygon,' 'square,' or 'buffer'.

#### 12.3.1.3 Parameter: buffer

Explanation: Applies only if 'cropway' = 'buffer'. Number parameter in kilometres (km) used to indicate the radius that will be considered to generate the circular areas around each coordinate/collection site or observed population.

#### 12.3.1.4 Parameter: shapefile

Explanation: Applies only if 'cropway' = 'polygon'. In this parameter, you must indicate the name of the shapefile file that must be in the WGS84 lat-long coordinate system and contain a single polygon that will be used as a template to crop. The shapefile is made up of four to seven files (including the essential files .shp, .dbf, and .shx) and must be located in the Pasaporte folder (CAPFITOGEN3/Pasaporte) in *local mode*. In *on server mode*, shapefile files must be previously uploaded to the User's Files and Results area, particularly in the 'uploads' folder. This way, the user only has to specify the name of the file in this parameter.

#### 12.3.1.5 Parameter: pasaporte

Explanation: Applies only if 'cropway' = 'square' or 'cropway' = 'buffer'. For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is



named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

### 12.3.1.6 Parameter: geoqual

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

### 12.3.1.7 Parameter: totalqual

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

### 12.3.1.8 Parameter: resol1

Explanation: Select the level of resolution of the ecogeographic information layers that you want to crop. In *local mode*, make sure the global layers ('world' folder inside 'rdatamaps') have the resolution indicated in this parameter (folders 1x1, 5x5, 10x10, or 20x20).

### 12.3.1.9 Parameter: uname

Explanation: For *local mode*, select one of the following names for the new set of cropped layers: 'user1', 'user2' or 'user3'. For *on server mode*, choose a simple name without spaces or special characters.

## 12.4. Results of rLayer

Once the analysis is finished and, rlayer will produce the new set of ecogeographic layers according to the configuration of the parameters previously detailed. The cropped layers will be saved in folders 'user1', 'user2' or 'user3' within the 'rdatamaps' folder (*local mode*) or will be available within the drop-down list of countries and sets of parameter layers as 'pais' (*on server mode*).



BANCO PORTUQUÊS  
DE  
CERNOPLASHA VEGETAL

National Workshop, Braga (Portugal), June 2015.



# 13 | Complementa Tool

## 13.1. *In situ* conservation of agrobiodiversity

*In situ* conservation is the only conservation strategy that allows organisms to naturally evolve and adapt to the constantly changing environmental conditions.

*In situ* conservation of plant genetic resources for food and agriculture occurs in two contexts:

- Protected areas, where the aim is to conserve a whole ecosystem in its natural state to ensure the long-term persistence of the species that occur within.
- On-farm and farming communities, where the aim is to conserve mainly local and/or traditional crop varieties in the natural agroecological conditions that allowed them to appear and survive; nevertheless, wild relatives can also be conserved.

The need to complement *ex situ* conservation efforts with *in situ* conservation activities and projects for the same target species is constantly stressed. The first Global Plan of Action (FAO, 1997) was one of the documents to clearly emphasize the strong need to *combine ex situ* conservation efforts along with *in situ* conservation plans and activities for a target species. In 2010, the adoption of the Second Global Plan of Action reiterates the need to combine both *in situ* and *ex situ* approaches. The plan highlights that ‘conservation and use strategies are most effective when they are complementary and well-coordinated. *In situ* conservation, *ex situ* conservation and sustainable use need to be fully integrated at all levels’ (FAO, 2012).

The priority activities for the *in situ* conservation of agrobiodiversity as listed in the Plan of Action are the following:

- Studying and inventorying plant genetic resources for food and agriculture.
- Supporting on-farm management and improvement of plant genetic resources for food and agriculture.
- Assisting farmers in disaster situations to restore crop systems that favour conservation.
- Promoting *in situ* conservation of crop wild relatives.

The objective of the first activity listed is to facilitate the development, implementation, and monitoring of conservation strategies, and to apply methodologies for inventorying PGRFA *in situ* and *ex situ*, including GIS and molecular markers. The objective of these studies is to identify, locate, inventory, and assess threats to agrobiodiversity, particularly from land use and climate change (FAO, 2012).

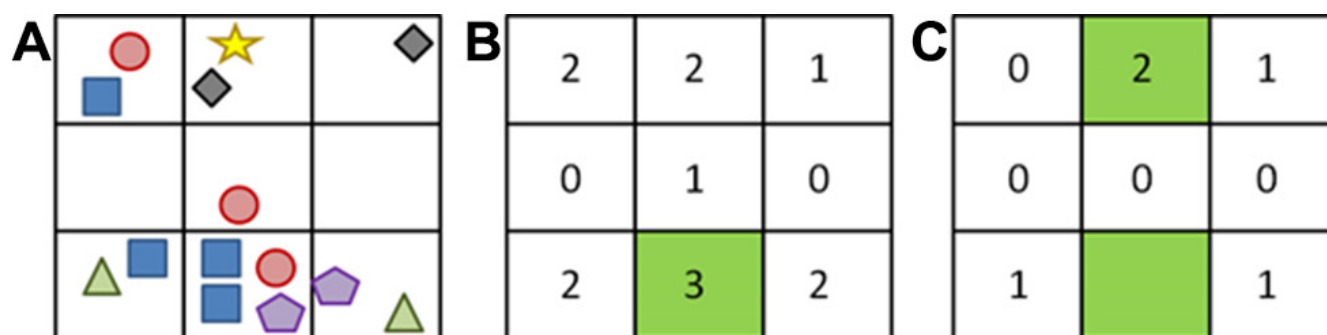
These specific objectives raise the need for adequate technical support to identify, assess and monitor sites designated for either on-farm or nature (protected areas) *in situ* conservation practices.

## 13.2 Concept of complementarity

Since the 1970s, the identification of suitable sites to be designated as protected areas for the conservation of biodiversity started as a methodological process. In the most important step of this process, a set of criteria is applied to prioritize areas for conservation. One of the criteria is the estimated value to conserve biological diversity, expressed either as species richness and/or diversity, rarity, or representativeness. Planification of such activities started to

move from a purely intuitive approach to an algorithmic one, and the criteria considered were automatically applied until all intuitive concepts had been eliminated. This change was facilitated by the development of modern computers available to researchers, and the development of Geographic Information Systems (GIS) (Justus & Sarkar, 2002).

The design of a network of protected areas for the conservation of biodiversity seeks to conserve the highest number of species in the smallest possible area (Kati *et al.*, 2004). A key feature in this network design is the use of the principle of complementarity, which ensures that areas chosen for inclusion in a reserve network complement those already selected. 'Complement' is defined in the dictionary as a 'thing that contributes extra features to something else in such a way as to improve or emphasize its quality'. This is the idea behind the concept of complementarity illustrated in Fig. 45.



**Figure 45.** How does the principle of complementarity work? A. Six different shapes representing different taxa scattered over nine cells. B. Number of different taxa occurring in each cell. The cell containing the highest number of taxa (3 taxa) is highlighted in green. C. The cell that contains the most different and complementary taxa to the cell selected in step B is highlighted in green and with the number '2' (two different and complementary taxa). The principle of complementarity is being applied in this last step C.

Complementarity was first used to design a network of protected areas in Tasmania (Australia) in 1983. It was then pointed out that non-iterative procedures for place selection were inadequate for the representation of the highest number of different taxa in the lowest possible number of sites. In the light of this, the use of richness as a criterion had to be replaced by an iterative procedure that considered complementarity within areas. Then, new algorithms were gradually developed for iterative procedures for place selection, based on the principle of complementarity (Justus & Sarkar, 2002).

In 1990, Rebelo and Siegfried (1990) developed an algorithm to select sites for the protection of the floral diversity in the Cape Floristic Region in South Africa. First, the algorithm starts with the selection of an area (grid square) with the highest number of endemic species (endemism and richness based), termed the 'primary core square'. The area is defined by overlapping the distribution of the different taxa over a grid of squares covering the whole territory. Those taxa overlapping the 'primary core square' are then removed from the overall distribution of taxa. Secondly, the grid square with the highest number of taxa (different from each other) and not present in the core squares is selected and designated as 'secondary core square'. Its taxa are then removed from the map, and the procedure is repeated. The al-

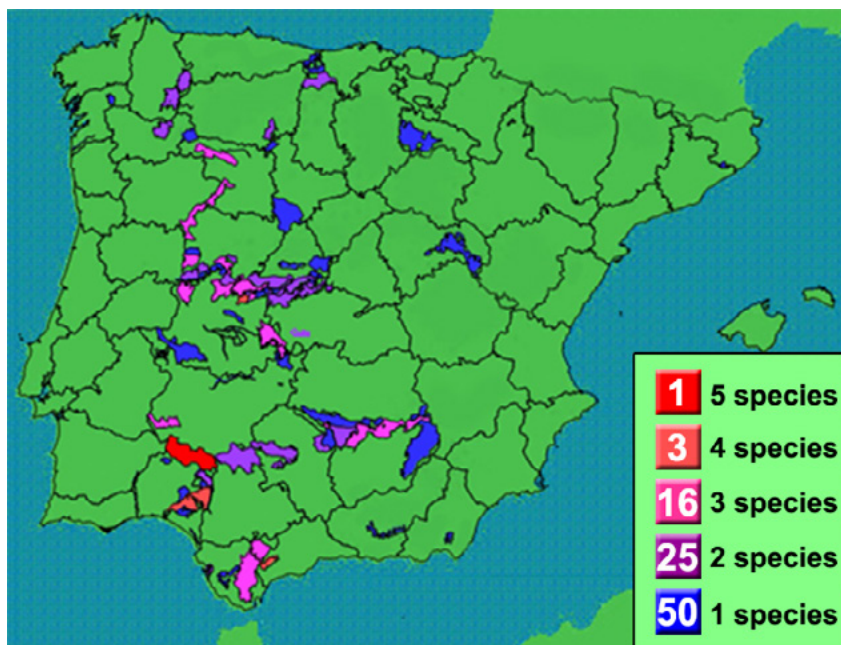
gorithm continues until a grid square is selected for the remaining taxon not covered in the previous step. Where ties occur (i.e., an equal number of complementary taxa in two or more grid squares), strategies such as random selection can be applied. Here, either grid squares with the highest species richness can be selected (considering both complementary and non-complementary taxa), or those grid squares with the highest number of endemic taxa instead. This iterative procedure is commonly known as Rebelo's complementarity algorithm, and it is widely used for identifying priority areas for conservation as well as for assessing already established networks of protected areas.

Rebelo's complementarity algorithm was designed to divide the whole territory into grid squares of the same size. However, when assessing current networks of protected areas, it can be implemented over areas with different shapes and sizes.

A complementarity analysis can also be used to define priority areas for germplasm collection for *ex situ* conservation, particularly useful for crop wild relatives. In this particular case, the designation of areas where taxa richness is high and complementary with each other will mean maximizing the collection of taxa in relatively small areas.

### 13.3 Concept of coverage

The concept of coverage as the protection of biodiversity provided by a network of protected areas is much simpler and more intuitive than the above concept of complementarity. Based on the number and richness of taxa protected, the analysis of the coverage seeks to quickly assess the degree of protection provided by a network of protected areas and each of the areas that constitute the network.



**Figure 46.** Map showing the coverage analysis for six *Lupinus* species in Spain. Sites of Community Importance were used as the network of protected areas to assess. The map shows those areas protecting the highest species richness.

Parra-Quijano *et al.* (2003) carried out a GIS assessment of the coverage provided by Sites of Community Importance (SCI) within the Natura 2000 network. In their study, they used six *Lupinus* species, including both cultivated and crop wild relatives that occur in the Iberian Peninsula and the Balearic Islands. The results demonstrated that one-third of the populations assessed are potentially protected under the network. Additionally, both the areas protecting the highest number of species and the species best protected were also determined. The network did not target crop wild relatives or cultivated species in its initial design which explains the low coverage obtained, and also the low match (1.47 %) between *Lupinus* species and the existing SCI.

## 13.4. How does Complementa work?

This tool for complementarity and coverage analysis requires the user to provide occurrence data of the taxa to be assessed. Occurrence data can be obtained from germplasm collections, georeferenced populations found in the literature, biodiversity databases, herbarium data, etc. Note that, regardless of the nature of the data, the required format will be 'passport FAO-Bioversity 2012' with two additional fields (optional and useful for specific complementarity analyses). However, unlike other tools, the Complementa 'passport' format has a reduced number of mandatory fields:

- ACCENUMB
- GENUS
- SPECIES
- SUBTAXA
- DECLATITUDE
- DECLONGITUDE

As the analysis becomes more complex, the table requires more information. The normal complementarity analysis run by Complementa targets a set of taxa (genera, species, and subspecies or varieties), identifying those sites where richness is highest for these taxa and complementary to each other.

### **13.4.1 Taxa or taxa-ecogeographic scenario analyses?**

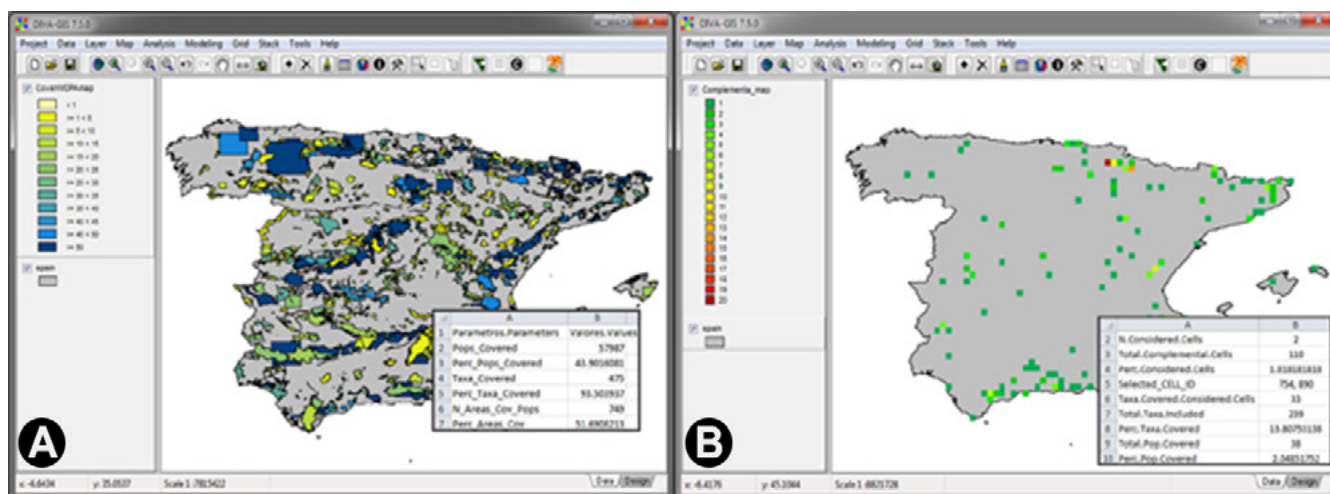
Complementa usually considers just the taxonomic aspects (fields GENUS, SPECIES, and SUBTAXA) of the 'passport' table to run the analysis. However, it can also run complementarity analyses combining both taxa and adaptive scenarios where each population occurs (i.e., ELC map categories). This last option requires the previous creation of an ELC map (ELCmapas tool). Also, occurrence sites need to be analysed using Representa tool before they are analysed with Complementa and the above mentioned ELC map.

One of the results obtained by the Representa tool is the table 'Genbank\_ELC.txt'. The table shows the field 'BGcat' which corresponds to an ELC map category that has been obtained according to the specific coordinates. To run the combined taxa/taxonomic-ecogeographic scenario analysis, it is necessary to transfer the information contained in the field 'BGcat' of the table 'Genbank\_ELC' to the field 'BG\_ELC' of the 'Complementa' passport table. The informa-

tion must be carefully transferred between tables by using a common ID field such as 'ACCENUMB'. Thus, Complementa will be able to combine the information and identify areas of high taxa richness and complementary to each other from a combined taxonomic-adaptive (abiotic) point of view.

### 13.4.2 Cells or areas?

Complementa can run complementarity analysis in a territory that has been divided into a square grid with same size cells or also in pre-existing areas. If one chooses to perform the analysis in a square grid territory, the recurring idea is to detect where the double condition (richness and complementarity) takes place. On the other hand, if the analysis is run on pre-existing protected areas, we seek to add extra value to some of the areas of the network for their capacity to protect as much biodiversity as possible in a smaller space. Note that the latter analysis will not necessarily show the areas within the target territory with the highest taxa richness and complementary to each other.



**Figure 47.** Maps resulting from a complementarity analysis. Analysis run: A. in a network of protected areas B. in a territory divided into cells.

### 13.4.3 Using Complementa for optimizing germplasm collections

If a complementarity analysis is used to identify collection sites, it is necessary to know the ecogeographic gaps within the collection. First, it will be necessary to use ELCmapas tool to obtain an ELC map, as well as Representa tool (with data from both the target germplasm collection and external sources). This first step is essential to identify high taxa richness areas complementary to each other from a taxonomic point of view, where populations occur that could potentially fill the ecogeographic gaps of the germplasm collection. In this kind of complementarity analysis, the user must consider as 'passport data' only the ecogeographic gaps in table 'Tabla\_Fuentes\_Externas\_clasificadas\_ExternalSourcesClassified.txt' which has been previously obtained with Representa. The information in the above table is to be carefully transferred to the 'passport' table used in Complementa, including all data from column 'Tipo\_falt/



Gap\_type' (Representa table) to column GAPATYPE (Complementa 'passport' table). It is important to realize now that the field GAPATYPE in the Complementa 'passport' table will only be filled when optimizing germplasm collections, and the only data to be inserted is either that of Table 1 or NA values.

Then, when filling the parameters in Complementa, the user will be required to answer questions on how to use the data in column GAPATYPE. For instance, what is the threshold under which a population coming from external sources data is considered a 'gap' or 'priority ecogeographic gap'? Does the user want to exclude NA values (not analysed with Representa) from the complementarity analysis?

#### **13.4.4 Using a network of protected areas from the World Database on Protected Areas or the user's own network of areas**

When the user decides to run a complementarity analysis using areas, Complementa allows two different types of networks of areas to be used:

##### **13.4.4.1 Network of protected areas from the World Database on Protected Areas - WDPA.**

It is a joint project of IUCN and UNEP and the most comprehensive global database on terrestrial and marine protected areas. ProtectedPlanet.net is the online interface (<http://www.protectedplanet.net/>). Given the fact that CAPFITOGEN3 was designed to facilitate the analysis to any user and minimal knowledge is required to operate the tools, all the relevant information on global coverage has been downloaded. The information is presented by countries, and the marine areas have been cut out from both terrestrial and marine protected areas. For Complementa to use this information, the following steps are to be followed:

- Visit the CAPFITOGEN website and follow the instructions on how to download the tools and additional information: <http://www.capfitogen.net/es/acceso/informacioncomplementa>
- This page displays a link to download data that have been adjusted and cropped from WDPA. Click on it.
- A list of '.RData' files for different countries and regions will come up. The user can download the file for the target country by clicking on it.
- Save the downloaded file in the 'wdpa' folder in the CAPFITOGEN directory (usually X:/CAPFITOGEN3/wdpa where X is the letter of the disk).
- When running Complementa tool, the user can now choose the target country or region in parameter 'pais'. Do not select a country or region for which a '.RData' file is missing in the 'wdpa' folder.

For Complementa in *on server mode*, all the .RData files for countries and regions are already loaded in the system. Therefore, in parameter 'pais', you should only select (from a drop-down list) the name of the country or region for which there are WDPA areas.

#### 13.4.4.2 User's own network of areas.

The user will provide a file with a GIS layer with polygons representing the protected areas to be assessed. For both modes (local and on server), the file should meet the following requirements:

- It must be a polygon 'shapefile' (vector data storage format).
- The 'shapefile' must consist of at least three mandatory files. For example, under the name 'AreasProt' at least the following files should be stored: 'AreasProt.shp', 'AreasProt.dbf' and 'AreasProt.shx'. Although not essential for Complementa tool, a complete 'shapefile' will also include the following files: 'AreasProt.sbn', 'AreasProt.sbx' and 'AreasProt.prj'. The 'shapefile' must be available in either the World Geodetic Coordinate System 84, better known as WGS84, or the geographic coordinate system.
- Before using the 'shapefile', double-check that the georeferencing of the layer is correct by overlapping it with the 'country boundaries shapefile'. This shapefile is located in the 'MapasApoyo\_BoundariesMaps' folder in the CAPFITOGEN directory. This step can be done using DIVA-GIS (<http://www.diva-gis.org>).
- A polygon shapefile is related to an attribute table displaying organized information (columns) for the different polygons (rows). The attribute table of the user's protected areas 'shapefile' should include an ID column to identify the polygons within. Every single polygon (protected area) must have a unique numeric, alphabetic, or alphanumeric ID code. Identify the name of the mentioned ID column using, for example, DIVA-GIS. The interface of Complementa will prompt the user to insert both the names of the 'shapefile' and the ID column.

Remember that for *local mode*:

- All the files included in the user's 'shapefile' must be copied to the 'wdpa' folder in the CAPFITOGEN directory.

And for *on server mode*:

- All the files included in the user's 'shapefile' must be uploaded to the User's Files and Results area, particularly in the 'uploads' folder.

#### 13.4.5 Coverage analysis

Complementa can run a coverage analysis with the same data used in the complementarity analysis of the areas (not available for 'cells' analysis). This analysis allows a more complete evaluation of the network of protected areas.

## 13.5. Using Complementa tool

Users can check the following studies where Complementa has been used, alone or in conjunction with other CAPFITOGEN tools: Phillips *et al.* (2016), García *et al.* (2017), Rubio-Teso *et al.* (2019), or Mponya *et al.* (2021). Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and Complementa tool has been selected, a series of parameters must be specified by the user.

### **13.5.1 Initial parameters defined by the user**

#### **13.5.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### **13.5.1.2 Parameter: pasaporte**

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area. For Complementa, this table must have additional fields such as BG\_ELC (ELC map category assigned by Representa) and GAPTYPE (type of ecogeographic gap or determined by Representa). It is recommended to have Excel files specially prepared for Complementa with the required columns, which can be found in <https://drive.google.com/drive/folders/1xCnllZgzW0uDeClDvcxbADv9H583xzpn?usp=sharing>.

#### **13.5.1.3 Parameter: geoqual**

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table ('pasaporte' parameter). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### **13.5.1.4 Parameter: totalqual**

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

**13.5.1.5 Parameter: gaptype**

Explanation: Applies only when running a complementarity analysis to prioritize collection sites (within an optimized design of germplasm collections). Select this option if there is data in the occurrence data table that can be used in the field GAPTYPE (produced by Representa) and you wish to use them to identify ecogeographic gaps within the collection.

**13.5.1.6 Parameter: gaptresh**

Explanation: Applies only if 'gaptype' has been set as TRUE (✓ in *on server mode*). Indicate the value of column GAPTYPE that will be used as a threshold to identify occurrence data as ecogeographic gaps within the collection. Occurrence data with GAPTYPE values below the threshold value will be included in the complementarity analysis.

**13.5.1.7 Parameter: gapna**

Explanation: Applies only if 'gaptype' has been set as TRUE (✓ in *on server mode*). When using the values in column GAPTYPE to identify ecogeographic gaps, indicate if NA values (occurrence data not assessed by the 'Representa' tool) are to be included (write 'include') in the analysis or not (write 'exclude').

**13.5.1.8 Parameter: duplicat**

Explanation: Indicate (TRUE=yes FALSE=no for *local mode*, ✓=yes □=no in *on server mode*) whether the records of the same GENUS/SPECIES/SUBTAXA will be deleted or not since they are spatially close to each other.

**13.5.1.9 Parameter: distdup**

Explanation: Applies only if 'duplicat' has been set as TRUE (✓ in *on server mode*). Determine the distance (in km) under which you consider that two observation or collection sites represent the same population (spatial duplicate). Sometimes two observers or collectors have registered two points that belong to the same population, but they do not know it. Only when the two points are mapped it is possible to determine that they belong to the same population by how close the points are to one another, and this is what distdup does. Parameter 'distdup' cannot be less than zero and it cannot be an extremely high number, since it would make all the points belong to the same population and the analysis would be performed for a single point, producing an error.

**13.5.1.10 Parameter: celdas**

Explanation: Select this option (TRUE in *local mode*, ✓ in *on server mode*) if you wish to run the complementarity analysis for a square grid covering all the occurrence distribution data.

### 13.5.1.11 Parameter: resol1

Explanation: Select the resolution level you wish to use to generate the complementarity map by cells. Note that 1x1 km offers greater resolution but requires greater computing capacity than 5x5 km; however, this is not as limiting a factor as it is for ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents, or continents.

### 13.5.1.12 Parameter: nceldas

Explanation: Applies only if 'celdas' has been set as TRUE (✓ in *on server mode*). The complementarity analysis for a square grid establishes a ranking of cells with high coverage values and that are complementary to each other; these cells also contain the highest number of different taxa within. Complementa can also generate some additional statistics for a specific subset of cells from the ranking. Choose the number of cells to be included in the additional analysis. The selection starts from the cells with the highest coverage values and continues down the ranking.

### 13.5.1.13 Parameter: areas

Explanation: Select this option (TRUE in *local mode*, ✓ in *on server mode*) if you wish to run the complementarity analysis on a set of pre-existing areas.

### 13.5.1.14 Parameter: WDPA

Explanation: Applies only if 'areas' has been set as TRUE (✓ in *on server mode*). Select this option if you wish to run the complementarity analysis using polygons from the World Database on Protected Areas (WDPA).

### 13.5.1.15 Parameter: pais

Explanation: Applies only if 'WDPA' has been set as TRUE (✓ in *on server mode*). Select the target country for which a complementarity analysis is required using areas from the WDPA. Double-check that for the target country there does exist a 'name of the country.RData' file in the 'wdpa' folder which is part of the set of folders that make up the CAPFITOGEN3 directory (local mode) that have been previously uploaded from <https://drive.google.com/drive/folders/1xCnllZgzW0uDeCldvcxbADv9H583xzpn?usp=sharing> In *on server mode*, just select the country or region from the drop-down list.

### 13.5.1.16 Parameter: propio

Explanation: Applies only if 'areas' has been set as TRUE (✓ in *on server mode*) and 'WDPA' as FALSE (☐ in *on server mode*). Select this option (TRUE in *local mode*, ✓ in *on server mode*) if you wish to run the complementarity analysis for areas/polygons using your own shapefile. Make sure that the shapefile has previously been copied and pasted into the 'wdpa' folder which is part of the set of folders that make up the CAPFITOGEN directory. Remember that a shapefile must be made up of at least three files stored under the same name and with the extensions shp, .dbf and

.shx. This shapefile should be in the geographic coordinate system (datum WGS84).

#### 13.5.1.17 Parameter: nombre

Explanation: Applies only if 'propio' has been set as TRUE (✓ in *on server mode*). Please Type the exact name of the shapefile with the polygons to be analysed. Do not include here the file extension or any other character such as dots. For example, if the shapefile comprises the following files: 'protectedareas.shp', 'protectedareas.dbf', or 'protectedareas.shx', do only type 'protectedareas' in the box.

#### 13.5.1.18 Parameter: campo

Explanation: Applies only if 'propio' has been set as TRUE (✓ in *on server mode*). Type the exact name as it appears in the polygon ID field of the shapefile. If unsure about the exact name, open the shapefile in DIVA-GIS and look it up in the table that accompanies the map of polygons.

#### 13.5.1.19 Parameter: nareas

Explanation: Applies only if 'areas' has been set as TRUE (✓ in *on server mode*). The complementarity analysis for either WDPA or the user's own network of areas ranks areas with a high coverage value. These areas have the highest number of different taxa and are complementary to each other. Complementa can also generate some additional statistics for a specific subset of areas from the ranking. Choose the number of areas to be included in the additional analysis. The selection starts from the areas with the highest coverage values and continues down the ranking.

#### 13.5.1.20 Parameter: coveran

Explanation: Applies only if 'areas' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode*, ✓ in *on server mode*) to run a coverage analysis (in addition to the complementarity analysis). Depending on the parameters that had been previously selected (i.e., 'WDPA' or 'propio'), the analysis can be performed for either WDPA, the user's own network of areas, or both.

#### 13.5.1.21 Parameter: niveltax

Explanation: Indicate here the taxonomic rank (from the following options: 'genus', 'species' or 'subtaxa') for which the complementarity analysis is required. This selection will influence the analysis regardless of the type (i.e., square grid, WDPA areas, or the user's own areas). Please note that the taxonomic rank information comes from the following fields in the occurrence data table: 'GENUS', 'SPECIES' or 'SUBTAXA'.

#### 13.5.1.22 Parameter: datanatax

Explanation: Select this option (TRUE in *local mode*, ✓ in *on server mode*) If you wish to include in the analysis NA values found in the selected field 'taxonomic rank' (i.e., 'GENUS', 'SPECIES' or 'SUBTAXA' according to parameter 'niveltax').

### 13.5.1.23 Parameter: `mapaelcf`

Explanation: Select this option (TRUE in *local mode*, ✓ in *on server mode*) if the occurrence data table contains information in the field 'BG\_ELC' and you want to combine it with the taxonomic information to perform a taxonomic-ecogeographic complementarity analysis.

### 13.5.1.24 Parameter: `mapaelc`

Explanation: Applies only if 'mapaelcf' has been set as TRUE (✓ in *on server mode*). Enter the name of the file (*local mode*) containing the ELC map generated by the ELCmapas tool or upload it to the server (*on server mode*). This map should be found in the CAPFITOGEN3/ELCmapas folder (*local mode*). The map should be in DIVAGIS format (.grd extension, exactly as generated by the ELCmapas tool) and its name should be entered with the file extension. Thus, if the name of the map is 'mapa\_elc\_spain', you should enter 'mapa\_elc\_spain.grd' (*local mode*) or search for this file (*on server mode*).

### 13.5.1.25 Parameter: `datanaelc`

Explanation: Applies only if 'mapaelcf' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode*, ✓ in *on server mode*) if you wish to include in the analysis NA values found in the field 'BG\_ELC'. By choosing this option, 'NA' will appear as a different category in the ELC map.

### 13.5.1.26 Parameter: `data0elc`

Explanation: Applies only if 'mapaelcf' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode*, ✓ in *on server mode*) if you wish to include in the complementarity analysis zero values found in the field 'BG\_ELC'. This means that ecogeographic category 0 of the ELC map will be considered as an adaptive scenario.

### 13.5.1.27 Parameter: `resultados` (only for *local mode*)

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 13.6. Results of Complementa

After defining all the parameters and paths (for *local mode*) that Complementa requires, click on the "Run" (*local mode* in RStudio) or "Start" (*on server mode*) buttons to start the analysis process of the tool.

After some time that may vary due to the type of analysis requested, Complementa will save the results in the path and folder specified in "resultados" (in *local mode*), or in the user's files and results area in *on server mode*.

Regardless of the mode used (local or on server), according to the requested analyses, the results will be produced in different folders:

- ‘AnalisisAreasProt\_ProtectedAreasAnalysis’ if the analysis was run for areas from a ‘shapefile’ supplied by the user.
- ‘AnalisisWDPA\_WDPAAnalysis’ if the analysis was run for WDPA areas.
- ‘AnalisisCeldas\_CellAnalysis’ if the analysis was run for cells.

Additionally, the user will find a folder named ‘CoverageAnalysis’ only if a coverage analysis (within the complementarity analysis for areas) was required. The results that can be found in each folder are described as follows.

### **13.6.1 ‘AnalisisAreasProt ProtectedAreasAnalysis’ folder**

**‘Tabla\_Estadisticas\_Stats\_Table\_Complementa.txt/.xls’:** This table shows some interesting data and descriptive statistics such as total number of complementary areas, number of areas considered (data provided by the user in parameter ‘nareas’), percentage of taxa covered by the complementary areas considered, etc.

**‘Tabla\_final\_Final\_Table\_Complementa.txt/.xls’:** In this table, areas with the highest taxa richness and that are complementary to each other are ranked from high to low. The field ‘areaID’ corresponds to the code given to each area in the ID field of the ‘shapefile’. On the other hand, if WDPA areas have been used for the complementarity analysis, the column name will be ‘wdpaid’ and the ID code will match that of the WDPA. In any case, the number of different and complementary taxa occurring in each area is next to the ID code.

**‘Tabla\_Pob\_en\_fuera\_AreasProt\_Table\_Pop\_in\_out\_ProtAreas.txt/.xls’:** This table shows the total list of occurrence sites analysed, also indicating which ones fall within or out of the network of protected areas. The code/field ACCENUMB in the table along with the taxonomic and ecogeographic information (only when the user, in the first place, provided this information in the ‘passport’ format) will help to identify the different areas. The fields with all the relevant information are on the right side of the table. Here, NA values indicate that the occurrence site does not belong to a network of protected areas, whereas cells showing any kind of ID or other information related to a particular protected area indicate that the occurrence site falls within the boundaries of the specified area.

**‘Tabla\_Pob\_en\_Top\_Areas\_Table\_Pop\_in\_Top\_Areas.txt/.xls’:** This table shows a list of occurrence sites falling within the selected complementary protected areas (i.e., those areas at the top of the table ‘Tabla\_final\_Final\_Table\_Complementa.txt’). The number of areas here corresponds to the number previously indicated by the user in parameter ‘nareas’.

**‘Tabla\_Table\_Base.txt/.xls’:** This table shows the basic inputs needed to obtain the complementarity analysis. It displays ‘1’ when the taxon is present in a protected area (identified with a code) or ‘0’ when absent.

**‘ComplementAreas.shp’:** This polygon shapefile, which can be opened/visualized in DIVA-GIS, shows the complementary protected areas of the table ‘Tabla\_final\_Final\_Table\_Complementa.txt’. The shapefile’s attribute table here



has two important fields: 'area ID' and 'N\_Taxa'. The 'area ID' field comes from the user's own shapefile and has the same original name. The field 'N\_Taxa' shows the number of complementary taxa occurring in each area. All this data matches that in table 'Tabla\_final\_Final\_Table\_Complementa.txt'.

**'final\_analyzed\_points.shp'**: This point shapefile, which can be opened/visualized in DIVA-GIS, shows the occurrence sites of the taxa analysed. The shapefile's attribute table has key fields to identify the occurrence sites (i.e., ACCENUMB from the 'passport' table), taxonomic information, coordinates, reference taxonomic or taxonomic-ecogeographic units for the complementarity analysis (field CHAIN), and information about the protected area where the population grows/occurs.

### **13.6.2 'AnálisisWDPA WDPAAanalysis' folder**

**'Tabla\_Estadísticas\_Stats\_Table\_Complementa.txt/.xls'**: Table that corresponds to the 'Tabla\_Estadísticas\_Stats\_Table\_Complementa.txt' table in the section above.

**'Tabla\_final\_Final\_Table\_Complementa.txt/.xls'**: Table that corresponds to the 'Tabla\_final\_Final\_Table\_Complementa.txt' table in the section above.

**'Tabla\_Pob\_en\_fuera\_WDPA\_Table\_Pop\_in\_out\_WDPA.txt/.xls'**: Table that contains information equivalent to that included in the 'Tabla\_Pob\_en\_fuera\_AreasProt\_Table\_Pop\_in\_out\_ProtAreas.txt' table in the section above.

**'Tabla\_Pob\_en\_Top\_WDPA\_Table\_Pop\_in\_Top\_WDPA.txt/.xls'**: Table that corresponds to the 'Tabla\_Pob\_en\_Top\_Areas\_Table\_Pop\_in\_Top\_Areas.txt' table in the section above.

**'Tabla\_Table\_Base.txt'**: Table that corresponds to the 'Tabla\_Table\_Base.txt/.xls' table in the section above.

**'ComplementaryWDPAmapping.shp'**: Corresponds to the map 'ComplementAreas.shp' described in the section above.

**'final\_analyzed\_points.shp'**: Corresponds to the map 'final\_analyzed\_points.shp' described in the section above.

**'TotalWDPAmapping.shp'**: A 'polygon shapefile' showing both terrestrial and marine (only their terrestrial part) protected areas from the WDPA for the target country/region selected in the complementarity analysis.

### **13.6.3 'AnálisisCeldas CellAnalysis' folder**

**'Datos\_por\_Celda\_Data\_by\_CELL.txt/.xls'**: This table lists all the grid cells showing occurrence sites for either the analysed taxa or the taxon-ecogeographic category combinations. The number of taxa or taxon-ecogeographic category combinations occurring in each cell is next to the field 'cell ID'.

**‘Datos\_por\_Taxa\_ELC\_Data\_by\_Taxa\_ELC.txt/.xls’:** This table lists all the taxa or taxon-ecogeographic category combinations occurring within a cell. The number of cells where taxa (or field ‘taxon- ecogeographic category combination’) occur is next to the field ‘taxa ID’.

**‘Tabla\_Estadisticas\_Stats\_Table\_Complementa.txt/.xls’:** This table is equivalent to the ‘Tabla\_Estadisticas\_Stats\_Table\_Complementa.txt’ table in the sections above. However, it refers to grid cells and not protected areas.

**‘Tabla\_final\_Final\_Table\_Complementa.txt/.xls’:** This table is equivalent to the ‘Tabla\_final\_Final\_Table\_Complementa.txt’ table in the sections above. However, it refers to grid cells and not protected areas.

**‘Tabla\_Pob\_en\_Top\_Celdas\_Table\_Pop\_in\_Top\_Cells.txt/.xls’:** This table is equivalent to the tables ‘Tabla\_Pob\_en\_Top\_WDPA\_Table\_Pop\_in\_Top\_WDPA.txt’ or ‘Tabla\_Pob\_en\_Top\_Areas\_Table\_Pop\_in\_Top\_Areas.txt/.xls’ in the sections above. However, it refers to grid cells and not protected areas.

**‘Tabla\_Table\_Base.txt’:** This table is equivalent to the ‘Tabla\_Table\_Base.txt/.xls’ table in the sections above. However, it refers to grid cells and not protected areas.

**‘Complementa\_map.grd/.tif’:** A ‘raster’ map that can be visualized in DIVA-GIS (.grd) or other software (.tif). It shows the final result of the complementarity analysis by cells contained in the table ‘Tabla\_final\_Final\_Table\_Complementa.txt/xls’.

**‘final\_analyzed\_points.shp’:** Map that corresponds to the map ‘final\_analyzed\_points.shp’ described in the sections above.

#### **13.6.4 “CoverageAnalysis” folder**

This folder can be found within the folders ‘AnalisisAreasProt\_ProtectedAreasAnalysis’ or ‘AnalisisWDPA\_WDPAAnalysis’ if the user decided to run a coverage analysis. The user will find the same type of tables and maps in both folders regardless of the type of network of areas analysed. Only two files have different names: a map and a table. Thus, the resulting files from the user’s own network of protected areas will display ‘AREAS’ or ‘AreasProt’, whereas those from the WDPA will display ‘WDPA’ or ‘AreasWDPA’. The following are examples of the resulting files that can be found when using the user’s own network of protected areas:

**‘Tabla\_AreasProt\_Cubriendo\_ProtAreas\_Covering\_Table.txt/.xls’:** This table shows a list of all the protected areas with their ID codes. The field ‘N\_pops’ indicates either the number of populations or occurrence sites from the passport table that occurs within an area. The field ‘N\_Diff\_Taxa’ indicates either the different number of taxa or taxon-ecogeographic category combination that occurs within an area.

**‘Tabla\_Estadisticas\_Cobertura\_Stats\_Table\_Coverage.txt/.xls’:** This table shows some simple statistics related to the coverage provided by the network of protected areas over the target taxa, such as total number of populations

(Pops\_Covered) or taxa (Taxa\_covered) covered and their percentages (Perc\_Pops\_Covered and Perc\_Taxa\_Covered, respectively), or number of areas covering populations (N\_Areas\_Cov\_Pops) and their percentages (Perc\_Areas\_Cov).

**'Tabla\_Taxa\_Cubiertos\_Table\_Taxa\_Covered.txt/.xls'**: This table corresponds to a list similar to that included in 'Tabla\_AreasProt\_Cubriendo\_ProtAreas\_Covering\_Table.txt' but focusing on taxa (or the taxon-ecogeographic category combination). It shows the following fields: an ID field for either taxa or taxon-ecogeographic category combinations (Taxa\_TaxaELC), total number of populations per taxon (Total\_N\_Pops), number of populations covered for this taxon by any protected area (Pops\_in\_AREAS) and percentage of populations from this taxon falling within a protected area (Perc\_Pops\_in).

**'CoverAREASmap.shp'**: A polygon shapefile, which can be opened in DIVA-GIS, showing in its attribute table all the available information related to the protected areas. Two key fields are found on the right side are: 'N\_pops' and 'N\_Diff\_Tax'. For each area, the first field shows the number of populations belonging to it, and the second one indicates either the number of taxa or different taxon-ecogeographic category combinations in each area.

## 13.7. References

FAO. 1997. Plan de Acción Mundial para la Conservación y Utilización Sostenible de los Recursos Fitogenéticos para la Alimentación y la Agricultura y la Declaración de Leipzig. Rome, Italy. 64p.

FAO. 2012. Segundo Plan de Acción Mundial para los Recursos Fitogenéticos para la Alimentación y la Agricultura. Rome, Italy. 104p.

García, R. M., Parra-Quijano, M., Iriondo, J. M. 2017. A multispecies collecting strategy for crop wild relatives based on complementary areas with a high density of ecogeographical gaps. *Crop Science* 57(3): 1059-1069.

Justus, J., Sarkar, S. 2002. The principle of complementarity in the design of reserve networks to conserve biodiversity: a preliminary history. *Journal of Biosciences*, 27(4): 421-435.

Kati, V., Devillers, P., Dufrêne, M., Legakis, A., Vokou, D., Lebrun, P. 2004. Hotspots, complementarity or representativeness? Designing optimal small-scale reserves for biodiversity conservation. *Biological Conservation* 120(4): 471-480.

Mponya, N. K., Chanyenga, T., Brehm, J. M., Maxted, N. 2020. *In situ* and *ex situ* conservation gap analyses of crop wild relatives from Malawi. *Genetic Resources and Crop Evolution* 68: 759-771.

Rebelo, A. G., Siegfried, W. R. 1990. Protection of fynbos vegetation: ideal and real-world options. *Biological Conservation*, 54(1): 15-31.

Parra-Quijano, M., Draper, D., Iriondo, J. 2003. Assessing *in situ* conservation of *Lupinus* spp. In Spain through GIS.

Crop Wild Relative 1: 8-9.

Phillips, J., Asdal, Å., Magos Brehm, J., Rasmussen, M., Maxted, N. 2016. *In situ* and *ex situ* diversity analysis of priority crop wild relatives in Norway. *Diversity and Distributions*, 22(11): 1112-1126.

Rubio Teso, M. L., Iriondo, J. M. 2019. *In situ* Conservation Assessment of Forage and Fodder CWR in Spain Using Phytosociological Associations. *Sustainability* 11(21): 5882.



# ESTACIÓN EXPERIMENTAL CENTRAL DE LA AMAZONÍA



National Workshop, Joya de los Sachas (Ecuador), December 2015.



# 14 | Bfuture Tool

## 14.1. Bioclimatic information (layers) and its future projections

Projected climate information (precipitation and temperature) plays a key role in studying the adaptation potential of a species to climate change scenarios. It allows, if necessary, adopting measures to protect species.

Species distribution models use environmental information, such as climate (i.e., independent climate variables or predictors), as GIS layers to predict species distributions. Any change in the environmental variables could potentially lead to a change in the future species distribution. Models for predicting climate change scenarios, mainly due to greenhouse gas emissions, can produce GIS layers to help us predict our target species' distribution.

Fortunately, researchers have created those GIS layers and made them available, free of charge, for climate change impact studies. Such is the case of the WorldClim Project (<http://www.worldclim.org>) under Robert Hijman's direction. Using global climate models (or general circulation models) under different spatial-temporal emissions scenarios, GIS layers showing future predictions have been created based on current information.

According to Intergovernmental Panel on Climate Change (IPCC, [https://www.ipcc-data.org/guidelines/pages/gcm\\_guide.html](https://www.ipcc-data.org/guidelines/pages/gcm_guide.html)), General Circulation Models (GCMs) representing physical processes in the atmosphere, ocean, cryosphere, and land surface, simulate the response of the global climate system to increasing greenhouse gas concentrations. Table 2 shows the GCMs available at the WorldClim website that correspond to those of the CMIP5 Project (Coupled Model Intercomparison Project phase 5).

With the release of WorldClim 2.1 (current version and available since 2017), GCMs have been projected again for the CMIP6 version. The new GCMs CMIP6 available on the WorldClim website are BCC-CSM2-MR, CNRM-CM6-1, CNRM-ESM2-1, CanESM5, GFDL-ESM4, IPSL-CM6A-LR, MIROC-ES2L, MIROC6, and MRI-ESM2-0.

In 2014, the IPCC selected four Representative Concentration Pathways (RCPs) defined by their total radiative forcing (cumulative measure of human emissions of greenhouse gas from all sources expressed in Watts per square meter) pathway and level by 2100. The four RCPs were chosen to represent a broad range of climate outcomes based on a literature review. For example, RCP2.6 represents the most optimistic scenario with a radiative forcing of 2.6 W/m<sup>2</sup>. RCP4.5 and RCP6.0 correspond to moderate scenarios, and RCP8.5 would be the worst-case scenario. For each RCP, the predicted increase in the global average temperature between 2081-2100 compared to 1986-2005 will range as follows: RCP2.6 from 0.3 to 1.7°C; RCP4.5 from 1.1 to 2.6°C; RCP6.0 from 1.4 to 3.1°C, and RCP8.5 from 2.6 to 4.8°C (IPCC, 2013). Ranges are determined by the CMIP5 modelling.



**Table 2.** Global climate models (general circulation models - GCMs) available at WorldClim (<http://www.worldclim.org>) for CMIP5 as GIS layers.

Model (code)	Institution or modelling centre	Terms of use
ACCESS1-0	CSIRO (Commonwealth Scientific and Industrial Research Organisation, Australia) and BOM (Bureau of Meteorology, Australia)	Unrestricted
BCC-CSM1-1	Beijing Climate Center, China Meteorological Administration	Unrestricted
CCSM4	Community Climate System Model, version 4	Unrestricted
CESM1-CAM5-1-FV2	National Science Foundation, Department of Energy, National Center for Atmospheric Research	Unrestricted
CNRM-CM5	Centre National de Recherches Meteorologiques / Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique	Unrestricted
GFDL-CM3	Geophysical Fluid Dynamics Laboratory	Unrestricted
GFDL-ESM2G		
GISS-E2-R	NASA Goddard Institute for Space Studies	Unrestricted
HadGEM2-AO	Met Office Hadley Centre (additional HadGEM2-ES realizations contributed by Instituto Nacional de Pesquisas Espaciais)	Unrestricted
HadGEM2-CC		
HadGEM2-ES		
INMCM4	Institute for Numerical Mathematics	Unrestricted
IPSL-CM5A-LR	Institut Pierre-Simon Laplace	Unrestricted
MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	Non-commercial only
MIROC-ESM		
MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	Non-commercial only
MPI-ESM-LR	Max Planck Institute for Meteorology (MPI-M)	Unrestricted
MRI-CGCM3	Meteorological Research Institute	Non-commercial only
NorESM1-M	Norwegian Climate Centre	Unrestricted

Source: Coupled Model Intercomparison Project website

WorldClim also adjusts RCPs for its new version (2.1) projected on GCM CMIP6. The concept of Shared Socioeconomic Pathways (SSP) is introduced given the fact that RCPs have been in use for 15 years and climate change scenarios have changed since then. Each of the nine SSPs that have been set up represents a different socioeconomic projection and political environment for the future. However, five SSPs have been prioritized in the Sixth Assessment Report - AR6 of the Intergovernmental Panel on Climate Change (IPCC), four of which are projected in WorldClim:

SSP1-2.6 (SSP126) of the first SSP 'sustainability' or low mitigation and adaptation challenges, where a radiative forcing of 2.6 W/m<sup>2</sup> is assumed for 2100; SSP2-4.5 (SSP245) of the SSP 'middle of the road' or intermediate mitigation and adaptation challenges, where a radiative forcing of 4.5 W/m<sup>2</sup> is assumed for 2100; SSP3-7.0 (SSP370) of the SSP 'fragmentation/regional rivalry' or high mitigation and adaptation challenges, where a radiative forcing of 7 W/m<sup>2</sup> is assumed for 2100, and finally SSP5-8.5 (SSP585) of the SSP 'conventional/fossil-fuelled development' or high mitigation and low adaptation challenges, where a radiative forcing of 8.5 W/m<sup>2</sup> is assumed for 2100 (Kebede et al., 2018; Meinshausen et al., 2020).

Future projections also vary according to the period to be estimated. As an example, under constant greenhouse gas emissions, the temperature increase will differ from the year 2050 to 2100. In the case of WorldClim, the projections available in version 2.1 are: 2030 (2021-2040), 2050 (2041-2060), 2070 (2061-2080) and 2090 (2081-2100).

## 14.2. How does Bfuture work?

In *local mode*, Bfuture provides the user with future climate information as GIS layers ready to be used in other tools, such as Modela. Two options are available to obtain this information:

First option: The user can download all the necessary data from the WorldClim website and save the downloaded zipped folder (.zip) into the 'rdatamapsf' folder. The tool itself will then unzip and convert the WorldClim format into the required CAPFITOGEN format. Optionally, the user can crop the layers to either the occurrence sites'/collection sites' distribution or the country/region boundaries. Please note that an internet connection is required to download the data from WorldClim but not to run Bfuture tool.

Second option: Bfuture tool can download all the predicted climate information from the WorldClim website. Internet connection is required throughout the whole process. As in the option above, Bfuture will then adapt the content to the required CAPFITOGEN format. Optionally, the user can also choose to crop the layers as needed. Note that regardless of the option chosen, the coverage/extent of the layers downloaded is global.

In *on server mode*, Bfuture only works by transferring files from WorldClim to the CAPFITOGEN3 server. Therefore, if the user considers using this mode, it is not necessary to download the layers from the WorldClim website.

In any case, before the using Bfuture tool, it is strongly advised to decide on the GCM, RCP scenario, and period to be used. To help make the choice, it is recommended to visit the WorldClim website (<http://www.worldclim.org>) and check similar studies on the impact of climate change on biodiversity in which this kind of information layers have

also been used. The user must define the kind of information or layers needed. The layers are organized into four groups: 'bioclimatic\_indices' that correspond to the 19 'bioclim' variables; 'monthly\_tot\_prec' or the 12 variables related to monthly total precipitation; 'monthly\_min\_temp' or the 12 variables related to monthly minimum temperature, and 'monthly\_max\_temp' or the 12 variables related to monthly maximum temperature. Bfuture allows the user to select either several sets of climate variables or all of them at once. It is important to highlight that WorldClim does not provide layers of future average monthly temperatures. Therefore, those variables will not be available for predictive modelling in tools such as Bfuture or Modela.

The user is also asked to define the cell size (spatial resolution) of the layers to download. WorldClim provides the following four resolutions: ~1x1, 5x5, 10x10, and 20x20 km (30 arcseconds, 2.5, 5 and 10 arcminutes, respectively). These resolutions are available for WorldClim version 1.0 and CMIP5. For version 2.1, only resolutions 5x5, 10x10, and 20x20 km (2.5, 5, and 10 arcminutes, respectively) are available as of January 2021. It is expected that the highest resolution (1x1 km or 30 arcseconds) will be available for WorldClim 2.1 and CMIP6 soon.

Once all five parameters have been defined (GCM, SSP, period, set of variables, and resolution) you have the necessary elements to use Bfuture both in *local mode* and *on server mode*.

Please note that the processes of downloading or transferring information layers using Bfuture or direct download from WorldClim can be time-consuming, depending on the internet speed. They will also require enough space in your hard drive (*local mode*), especially if downloading high-resolution layers (e.g., 1x1 km). As an example, for any GCM a set of bioclim layers is about 3.4 GB. For a fast and easy download, both high-speed internet and enough free space in the hard drive (*local mode*) are required. Bear in mind that the 3.4 GB file must be unzipped afterward, duplicating its size. Therefore, about 7 GB altogether should be free in the hard drive for some time in the process. If Bfuture tool has been used to download the information from WorldClim, through either *local mode* or *on server mode*, the .zip file will be automatically deleted once the process is completed.

If the user wants to crop global layers using the country boundary (parameter `croplayer=TRUE` and `paiscrop=TRUE`), ecogeographic layers for the country that will be used as a template must be available inside a folder with the selected resolution. This folder must be inside another folder with the name of the country which, in turn, must be inside the 'rdatamaps' folder. For instance, if the user wants to clip a 5x5 km resolution global layer of future predictions to Portugal's boundaries, '.tif' files (layers) should have already been saved in the following path: CAPFITOGEN/rdatamaps/portugal/5x5/.

The cropping process for occurrence/accessions data is very similar to that completed by rLayer tool.

### 14.3. Using Bfuture tool

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and Bfuture tool has been selected, the user should specify a series of parameters.

### **14.3.1 Initial parameters defined by the user**

#### **14.3.1.1 Parameter: ruta (only for local mode)**

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN, C:/CAPFITOGEN, D:/MisHerramientas/CAPFITOGEN, etc.

#### **14.3.1.2 Parameter: local (only for local mode)**

Explanation: Select this option (TRUE) if the predicted bioclimatic information at a global scale has been previously downloaded. The .zip file should be available in the 'rdatamaps' folder in the set of folders CAPFITOGEN3. If the box is left unchecked, future bioclimatic information will automatically be downloaded from the WorldClim website, so an internet connection is required.

#### **14.3.1.3 Parameter: resol1**

Explanation: Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution but requires greater computing capacity than 5x5 km; however, this is not as limiting a factor as it is for ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents, or continents. If the file of future bioclimatic data was already downloaded from the WorldClim website (*local mode* only), note that the resol1 value corresponds to the term that appears after wc2.1\_, at the beginning of the file name. For example, if your file's name is wc2.1\_10m\_tmin\_BCC-CSM2-MR\_ssp126\_2021-2040.zip, the resolution would be 10m, i.e., 10 minutes or resol1 = 'celdas 20x20 km aprox (10 arc-min)' or 20x20 km cell size. Make sure the resol1 value selected here matches the resolution value of the .zip file (10m= 'celdas 20x20 km aprox (10 arc-min)') so that Bfuture can properly extract and adjust the information contained in the .zip file.

#### **14.3.1.4 Parameter: ssp**

Explanation: Select the shared socioeconomic pathway (SSP) scenario for which you want to download or transfer information (*local mode* or *on server mode*) or for which you have already downloaded a .zip file from the WorldClim website (*only for local mode*). There are four SSPs available for CMIP6 in WorldClim: 126 (ssp1, 2.6), 245 (ssp2, 4.5), 370 (ssp3, 7.0), and 585 (ssp5, 8.5). If the user has already downloaded the future bioclimatic data file from the WorldClim website, note that the ssp value corresponds to the number that appears after the letters ssp. For example, if your file's name is wc2.1\_10m\_tmin\_BCC-CSM2-MR\_ssp126\_2021-2040.zip, the ssp would be 126. Make sure the ssp value selected here matches the ssp value of the .zip file so that Bfuture can properly extract and adjust the information contained in the .zip file.

#### 14.3.1.4 Parameter: gcm

Select the global climate model (GCM also known as general circulation models) for which the user wants to download the information or has already downloaded it as a .zip file from the WorldClim website. If the future bioclimatic data file was already downloaded from the WorldClim website (only for *local mode*), note that the gcm value corresponds to the term that follows the type of layers requested in the name of the .zip file. For example, if your file's name is `wc2.1_10m_tmin_BCC-CSM2-MR_ssp126_2021-2040.zip`, the gcm name comes after the layers requested (parameter 'varset', in this case, minimum temperatures or `tmin_`), i.e., 'BCC-CSM2-MR'. Make sure the option gcm selected here matches that of the .zip file so that Bfuture can properly extract and adjust the information contained in the .zip file.

#### 14.3.1.5 Parameter: proy

Explanation: Select the period for the future bioclimatic projected data you want to download or have already downloaded from the WorldClim website. Option 50 represents the value/year 2050 (average for 2041-2060) and option 70 represents the year 2070 (average for 2061-2080). If the user already downloaded the .zip file with future bioclimatic data from the WorldClim website (only for *local mode*), note that the 'proy' value corresponds to the interval that appears at the end of the name of the .zip file that was downloaded from WorldClim. Using the same example as before, if the name of your .zip file is `wc2.1_10m_tmin_BCC-CSM2-MR_ssp126_2021-2040.zip`, then the proy value would be 2030 since the interval of the file is '2021-2040'. Make sure the proy value selected here matches the time interval value in the .zip file so that Bfuture can extract, organize, and adjust to CAPFITOGEN format all the information contained in the .zip file.

#### 14.3.1.6 Parameter: varset

Explanation: Select the type of projected bioclimatic variables that you want to transfer (*local mode* or *on server mode*) or that you have already downloaded as a .zip file from the WorldClim website. If you have already downloaded the future bioclimatic data file from the WorldClim website (only for *local mode*), note that option 'varset' corresponds to the following term that defines the resolution. Using the previous example, if your file's name is `wc2.1_10m_tmin_BCC-CSM2-MR_ssp126_2021-2040.zip`, the varset value would be 'tmin' which corresponds to a minimum temperature ('monthly\_min\_temp'). Make sure that option 'varset' selected here matches the varset code of the .zip file so that Bfuture can properly extract and adjust the information contained in the .zip file.

#### 14.3.1.7 Parameter: croplayer

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you wish to crop the information (layers) either to the country boundary or to the extent area of the species distribution data. Option FALSE (*local mode*) or leaving the box unchecked ( in *on server mode*) indicates that the information downloaded keeps its original global coverage. Note that high spatial resolution such as 'resol1' with values of 'Celdas 1x1 km aprox (30 arc-seg)' (1x1 km cell size) or 'Celdas 5x5 km aprox (2.5 arc-min)' (5x5 km cell size) requires a lot of space in your hard drive (local mode) or in the User's Files and Results area (on server mode) to keep layers with global coverage.

#### 14.3.1.8 Parameter: paiscrop

Explanation: Applies only if 'croplayer' has been set as TRUE (✓ in *on server mode*). Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to crop the downloaded layers to the country/region boundary. To leave the box unchecked indicates that the layers will be cropped using your occurrence or distribution data (CAPFITOGEN passport format).

#### 14.3.1.9 Parameter: pais

Explanation: Applies only if 'paiscrop' has been set as TRUE (✓ in *on server mode*). Select the country/region that will define the crop of the layers. Note: Check the table in the following the link <http://www.capfitogen.net/en/tools/coverage/> to make sure that the spatial resolution selected in parameter 'resol1' is available for the country/region selected here. A mismatch between 'pais' and 'resol1' (according to the table) will lead to an error. Additionally, current data for both country/region and resolution should be available in the 'rdatamaps' folder (only for *local mode*).

#### 14.3.1.10 Parameter: pasaporte

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area. The geographic extent of the occurrence data included in this table will be used to crop the GIS layers downloaded.

#### 14.3.1.11 Parameter: geoqual

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

#### 14.3.1.12 Parameter: totalqual

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0

(zero quality) to 100 (maximum quality).

#### 14.3.1.13 Parameter: unname

Explanation: Applies only if 'croplayer' has been set as TRUE (✓ in *on server mode*) and 'paiscrop' has been set as FALSE (□ in *on server mode*). Select one of the following names for the new set of cropped layers by the distribution data (using the table indicated in 'pasaporte').

## 14.4. Results of Bfuture

Once Bfuture finishes the processes of data transfer (if data was not downloaded directly from WorldClim), decompression, adaptation to the CAPFITOGEN format, and (optionally) layer cropping, the tool provides the user with future projected layers (results) ready to be used. These layers can be found in the 'rdatamapsf' folder (*local mode*) or in the User's Files and Results area (*on server mode*). The 'Parametros.Parameters.Bfuture.txt' file can also be found within the same folder. This file contains the list of parameters used by the Bfuture tool. Depending on the cropping option used, the user will also find one of the following folders:

- a) Not cropped: a folder named 'world'
- b) Cropped to the distribution of occurrence data/accessions: a single folder named either 'user1', 'user2', or 'user3' (*local mode*) or the name that has been selected (*on server mode*) according to what was specified in parameter 'unname'.
- c) Cropped to country/region boundary: a folder with the name of the country/region.

Inside any of these folders above (i.e., 'world', 'userX' or 'country/region'), the user will find another folder named using the following combination separated by underscore symbols:

- a) GCM name
- b) SSP code
- c) Period of the future projection.

Within this folder you will find another folder whose name indicates the spatial resolution (parameter 'resol1'). In this case, the name of the folder could be one of the following: '1x1', '5x5', '10x10', and '20x20'. Here, the user will find the final future climate layers (.tif) already adapted and cropped (optional). The following is an example of the path where a future climate layer can be found. Using the example of Ecuador, choosing the MIROC6 model, SSP5 8.5, period 2090, a set of 'bioclim' variables and a 5x5 km resolution, the path to the first future climate layer (i.e., 'bio 1' or average annual temperature) would be: 'X:\CAPFITOGEN3\rdatamapsf\ecuador\MIROC6\_585\_2090\5x5\bio\_1.tif'

## 14.5. References

Kebede, A. S., Nicholls, R. J., Allan, A., Arto, I., Cazcarro, I., Fernandes, J. A., Hill, C.T., Hutton, C.W., Kay, S., Lázár, A.N., Macadam, I., Palmer, M., Suckall, N., Tompkins, E.L., Vincent, K., Whitehead, P. W. (2018). Applying the global RCP–SSP–SPA scenario framework at sub-national scale: A multi-scale and participatory scenario approach. *Science of the Total Environment* 635: 659-672.

IPCC, 2013. Summary for Policymakers. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. In: Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Meinshausen, M., Nicholls, Z. R., Lewis, J., Gidden, M. J., Vogel, E., Freund, M., Beyerle, U., Gessner, C., Nauels, A., Bauer, N., Canadell, J.G., Daniel, J.S., John, A., Krummel, P.B., Luderer, G., Meinshausen, N., Montzka, S.A., Rayner, P.J., Reimann, S., Smith, S.J., van den Berg, M., Velders, G.J.M., Vollmer, M.K., Wang, R. H. 2020. The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500. *Geoscientific Model Development* 13(8): 3571-3605.







**National Workshop, Tlalnepantla (Mexico), May 2015.**



# 15 | Modela Tool

## 15.1. Species Distribution Models (SDM)

Species Distribution Models (SDM) predict the species' occurrence or absence within a spatial context. Under various hypotheses, predictive modelling seeks to determine patterns between a species' absence or presence and environmental factors (Guisan & Zimmermann, 2000). As such predictive models are related to geographical frames, the use of GIS to visualize the results remains the perfect tool in SDM.

In agrobiodiversity conservation, SDM have become a very useful tool in decision-making processes for their ability to predict a species' potential distribution (Parra-Quijano *et al.*, 2012b). They have been used to:

- Optimize germplasm collections (Parra-Quijano *et al.*, 2012c).
- Support the collection of germplasm with special traits (Jarvis *et al.*, 2005).
- Identify gaps in *ex situ* conservation (Ramírez *et al.*, 2010).
- Study the impact of climate change on CWR species distributions (Jarvis *et al.*, 2008)
- Identify suitable sites for the establishment of genetic reserves for CWR (Parra-Quijano *et al.*, 2012a)

Recently, these models have often been used to determine plant species extinction risk under climate change. The 'Training Manual on Spatial Analysis of Plant Diversity and Distribution' (Scheldeman & van Zonneveld, 2011) provides detailed information on SDMs and their application on agrobiodiversity conservation (pages 139 and 140). This manual can be downloaded here: <http://goo.gl/XSiflm>. The manual specifies three essential aspects to consider when using modelling techniques to obtain reliable results:

- a) In theory, SDMs should only be used for wild species as their populations show high site-specific adaptation. Plant domestication gradually reduces specific adaptation, broadening the original distribution range of the wild ancestor. As a result, species native to a continent can also thrive in others. However, SDMs can still generate reliable predictions for semi cultivated species (species not fully domesticated) or traditional local varieties that have been locally cultivated for centuries as both may show enough site-specific adaptation. In any case, this last consideration may be subject to debate. Users can consider basic mechanistic models such as EcoCrop for cultivated species (Ramírez- Villegas *et al.*, 2013).
- b) Species distribution modelling not only requires occurrence data (presence/absence) but also predictive abiotic environmental variables. Both, species distribution and abiotic variables will determine a pattern (the model itself) which, in turn, is projected on environmental GIS layers covering the spatial work frame. Therefore, environmental variables used in the modelling must be determinant in terms of species adaptation and distribution.
- c) No occurrence data should be included of species growing in human-modified environments specially designed for the survival of a species. For cultivated species, occurrence data should not come from plantations in which the natural environmental conditions have been significantly modified to allow sowing (permanent irrigation, greenhouses, chemical or physical soil alterations, etc.). For wild species or forms, occurrence data should not correspond to planted specimens in gardens, arboretums, repopulated sites (with non-local or alien germplasm) or plantations (not as a wild species but as a cultivated one). The accuracy of the resulting models and species distribution predictions will highly depend on the quality of the georeferenced data (both presence-only and presence-absence data), along with the reliability of the information contained in the predictors' data layers.

## 15.2. Selecting variables to model - SelecVIF Tool

The results of species distribution modelling processes depend on many factors, such as the quality of population presence/absence data (dependent variables), the appropriate use of modelling algorithms, or the way ecogeographic or environmental variables are used (Sillero & Barbosa, 2021). The set of ecogeographic variables (independent variables) and their ability to serve as predictors of the distribution due to abiotic adaptation of the target species constitute another important factor. CAPFITOGEN3 offers a tool that helps its users to identify variables of high importance when creating ecogeographic groups (via Random Forest). SelecVar can also filter variables linearly, discarding those that would only provide redundancy (via bivariate correlation determination). This way, variables that provide unique information (at least from a linear point of view) will only be included in subsequent processes such as the creation of ELC maps or species distribution modelling.

The tool called SelecVIF was included to provide more discriminatory analyses of variables for species distribution modelling processes. This tool determines the value of Variance Inflation Factor - VIF (for more details, see [https://en.wikipedia.org/wiki/Variance\\_inflation\\_factor](https://en.wikipedia.org/wiki/Variance_inflation_factor)) of a set of ecogeographic variables that the user wants to evaluate, detecting multicollinearity problems and, thus, filtering those variables that show higher VIF values. This process is applied in variable removal for regression techniques (Akinwande *et al.*, 2015) and it is also very common in species distribution modelling processes (Guisan *et al.*, 2002).

SelecVIF allows an automatic selection of a set of variables with low multicollinearity, by performing a step-by-step elimination of those whose VIF values exceed a user-defined threshold. The VIF values of the remaining variables are determined in each step again, and the process is repeated until the established condition is met (VIF values below the threshold). For this reason, after running Modela tool, SelecVIF can be used as a complement to establish a criterion of inclusion or exclusion of variables. However, the selection of variables can also be based on the use of SelecVar or both, selecVar and SelecVIF sequentially.

As a model is required to determine VIF, the parameters for SelecVIF and Modela are quite similar. In this chapter, the section that describes parameters also includes those required by SelecVIF.

## 15.3. Presence-only or presence-absence data

The models establish relationship patterns between species presence and absence data and the environment where it occurs. It is important to point out that models require data on both species' presence and absence. In agrobiodiversity, only presence data are available which corresponds to the surveyed populations/individuals.

For this reason, it is necessary to generate pseudo-absences, that is locations of assumed species absence instead of confirmed species absence. The model could now be obtained and assessed with both presence and pseudo-absence data. There exist algorithms that allow modelling using presence-only data; however, by generating pseudo-absence data one can use any available algorithm.

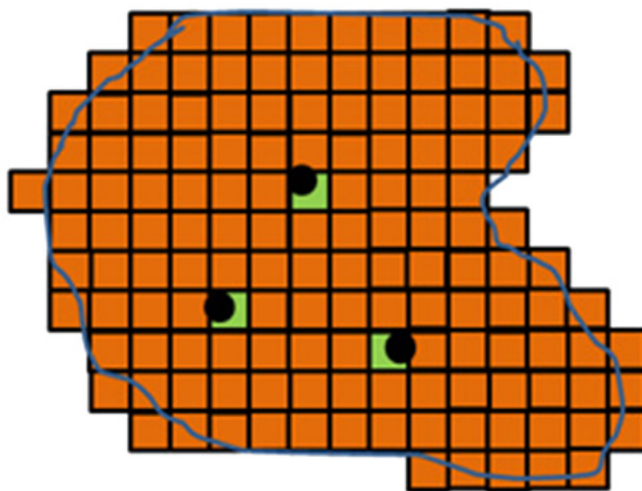
### **15.3.1 Methods to generate pseudo-absence data**

Since the introduction of pseudo-absence data instead of confirmed absence data, numerous techniques have been proposed to generate this type of data. The aim is to obtain more accurate modelling results. Several authors (Chefaoui & Lobo, 2008; Phillips *et al.*, 2009; Barbet-Massin *et al.*, 2012) carried out comparative studies to test the effect of different types of pseudo-absences for species modelling.

The spatial distribution of pseudo-absence data is an important factor to consider as it affects the results of the models, especially when the projection goes beyond the extent of the presence point data (VanDerWal *et al.*, 2009).

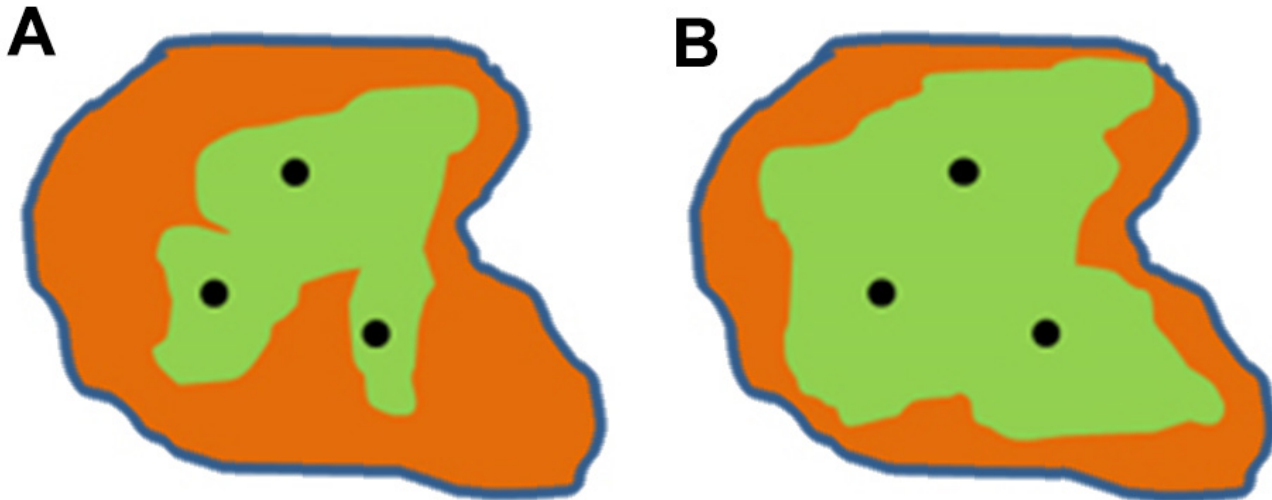
Modela provides four different ways of generating pseudo-absence data:

- a) **Random:** Pseudo-absence points are randomly selected from any cell of the study area not occupied by presence points. Hijmans and Elith (2015) describe this option as ‘background data’. According to the authors, background data establishes the environmental domain of the study, whilst presence data should establish the conditions under which a species is more likely to be present than on average. The authors consider ‘pseudo-absences’ as a relatively different concept in which absence data is guessed. They prefer the background concept since it requires fewer assumptions and has some coherent statistical methods for dealing with the overlap between presence and background points.



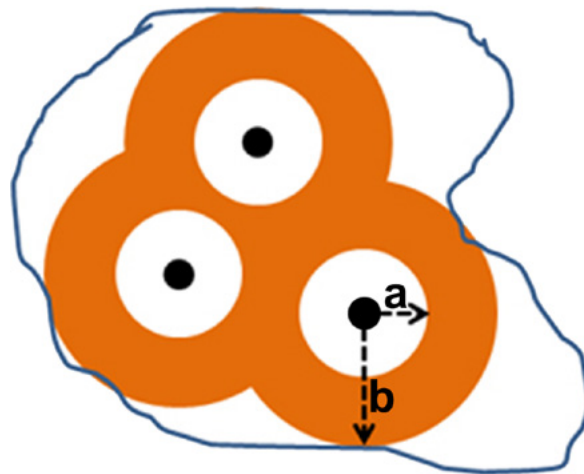
**Figure 48.** Random-type selection of pseudo-absence or “background” data. In the study area, pseudo-absence points will be selected from the orange cells that exclude presence data (black dots).

- b) **SRE (surface range envelop):** Pseudo-absence points will be selected from areas different from those of high probability of occurrence determined by surface range envelop models (i.e., BIOCLIM models). To determine such high probability, this option requires the definition of quantile values. The closer the quantile value is to 1, the envelope (i.e., suitable area for the species to occur) becomes larger reducing the area for selecting pseudo-absence points. This option can be misleading and overestimate the model predictions.



**Figure 49.** SRE-type selection of pseudo-absence points. Here, pseudo-absence points will be randomly selected within the orange area of low probability of occurrence determined by the Bioclim model (green areas correspond to a high probability of occurrence). In case A, a lower quantile compared to case B (quantile closer to 1) has been selected. An increase in the quantile value reduces the area for selection.

- c) **Disk:** This option draws a circle around each presence point. The distance to the presence point is selected by the user. To draw the disk, the user has to set both the minimum and maximum distance to the presence point. This option summarizes some other methods of generating pseudo-absence points at a certain distance from presence points suggested by Barber-Massin *et al.* (2012).



**Figure 50.** Disk-type selection of pseudo-absence points. Here, pseudo-absence points will be randomly selected within the orange disks. Such disks are drawn by setting both the minimum (*a*) and maximum (*b*) distance from the presence point.

- d) **ELC:** Modela tool introduces this novel approach for generating pseudo-absence points in/for SDMs. It is based on an ELC map previously created by ELCmapas tool for the target species (Chapter 6). These maps do not

provide probabilities but simply divide the territory into units (categories) of high environmental similarity. Then, by overlapping presence points within the ELC map, it is possible to obtain occurrence frequencies of the species for each category in the map. Similar to Representa's method, frequencies can be grouped in quartiles as follows:  $>0.75$  high,  $0.5-0.75$  mid-high,  $0.25-0.5$  mid-low, and  $<0.25$  low frequency. The group of categories where the species does not occur is also considered (null frequency). Based on this classification of the categories by frequencies, Modela tool allows the user to select the group(s) of categories for which pseudo-absence points are to be generated. Opposite to the SRE method, the ELC option prevents overestimations. If choosing this option, it is recommended that the ELC maps are of the same extent and cell size (resolution) to the ecogeographic layers used for the modelling.



**Figure 51.** ELC-type selection of pseudo-absence points. Here, pseudo-absence points are selected from the groups of ELC categories indicated by the user. Pseudo-absence points will be randomly selected from the green ('high frequency' group), blue ('mid-high' frequency group), pink ('mid-low' frequency group), yellow ('low' frequency group), or grey (null frequency) cells. Modela tool allows multiple group selection.

## 15.4. Algorithms for modelling

An algorithm for modelling is a set of mathematical operations for determining models or patterns between dependent and independent variables. The dependent variable being the presence-only or presence-absence data, and the independent variable being the ecogeographic predictors. There are numerous different algorithms for modelling potential species distributions. Guisan and Zimmermann (2000), Guisan and Thuiller (2005) or Elith and Leathwick (2009), amongst others, give a detailed explanation regarding the theory behind species distribution modelling, the available algorithms for modelling, the rules for algorithm selection, and the data needed in each type of algorithm (i.e., presence-absence or presence-only data; quantitative, semi-quantitative, or qualitative predictors).

For CAPFITOGEN users that wish to use Modela tool, it is recommended to do first a literature search on the available algorithms (parameters 'dismodel' and 'modelos'). The available algorithms in Modela are specified in Table 3. The



table also lists some useful references on the use of each algorithm in species distribution modelling.

Numerous references compare the efficiency between algorithms. They demonstrate that the appropriate choice of an algorithm depends on several factors, such as type of data, work frame, predictors' resolution, or the nature of the target species (see Guisan & Thuiller, 2005).

**Table 3.** Available algorithms for modelling in Modela tool. These algorithms are used as they are implemented in the R packages 'dismo' and 'biomod2'.

Algorithm code	R- package	Name	Reference to algorithm use in SDM
Domain	dismo	Domain	Carpenter <i>et al.</i> , 1993
Bioclim / SRE	dismo	Bioclim / Surface Range Envelope	Booth <i>et al.</i> , 2014
GLM	biomod2	Generalized Linear Model	Guisan <i>et al.</i> , 2002
GBM	biomod2	Generalized Boosting Model (Boosted Regression Trees)	Elith and Leathwick, 2008
GAM	biomod2	Generalized Additive Model	Guisan <i>et al.</i> , 2002
CTA	biomod2	Classification Tree Analysis	Thuiller <i>et al.</i> , 2003
ANN	biomod2	Artificial Neural Network	Özesmi and Özesmi, 1999
FDA	biomod2	Flexible Discriminant Analysis	Maiorano <i>et al.</i> , 2013
MARS	biomod2	Multiple Adaptive Regression Splines	Mateo <i>et al.</i> , 2010
RF	biomod2	Random Forest	Bradter <i>et al.</i> , 2013
Maxent	dismo y biomod2	Maximum Entropy	Elith <i>et al.</i> , 2011

## 15.5. Evaluation of models

Assessing the accuracy of predictive models is based on two aspects: reliability and discrimination ability. The reliability of predictions refers to how closely predicted probabilities match observed proportions of occurrence, and the discrimination ability defines how well models can discriminate between occupied and unoccupied areas (Pearce & Ferrier, 2000). A range of indices is used to evaluate both aspects. A number of these can only be applied to binary results or to continuous results that have been transformed into a binary solution by using a specific cut-off value, called a threshold. These indices are called ‘threshold-dependent’ indices and are based on the elements of the confusion table or matrix below (Liu *et al.*, 2009) (Table 4).

**Table 4.** Confusion table used to determine threshold-dependent indices.

		Prediction	
		Presence	Absence
Observation	Presence	a (true positive)	b (false positive)
	Absence	c (false negative)	d (true negative)

Accuracy measures (hereafter called ‘evaluators’ or ‘evaluation methods’) that use elements of the confusion table are used to assess the predictive ability of the model. The matrix can only be generated by dividing the data set into two different sets: ‘test set’ and ‘train set’. The former corresponds to the presence and absence data (or pseudo-absence data) that is not included in the modelling process; the latter is the presence and absence data (or pseudo-absence data) to be used to generate the model. In Modela, the user indicates the percentage of data to be used as ‘train set’ and ‘test set’ (parameters ‘datadiv’ or ‘datadiv2’).

Evaluators use elements of the confusion matrix, for example:

- Sensitivity =  $a / (a+c)$
- False negative fraction =  $c / (a+c)$
- Specificity =  $d / (b+d)$

The Relative Operating Characteristic (ROC) is an example of the relationship between evaluators. The ROC graphically describes the compromises that are made between the sensitivity and 1-specificity.

The following are the evaluators available in Modela:

- ROC (Relative Operating Characteristic), see Fawcett (2004).
- Cohen’s Kappa, see Cohen (1960) or Wood (2007).
- TSS (True skill statistic), see Allouche *et al.* (2006)
- FAR (False alarm ratio) applied and explained by Barnes *et al.* (2007), Barnes *et al.* (2009), and Roeber *et al.* (2009) regarding meteorology studies.

- e) SR (Success ratio) equals 1-FAR, see Roeber *et al.* (2009)
- f) ACCURACY (fraction correct), see explanation at <http://goo.gl/jEL02V>
- g) BIAS (Bias score or frequency bias), see Roeber *et al.* (2009)
- h) POD (Probability of detection), see Roeber *et al.* (2009)
- i) CSI (Critical success index), see Roeber *et al.* (2009)
- j) ETS (Equitable threat score), see explanation at <http://goo.gl/jEL02V>

Follow the link <http://goo.gl/jEL02V> for descriptions of the evaluators, along with examples of their use in assessing climate predictions. For the appropriate choice of evaluation methods, it is highly recommended to review the available literature on the evaluation methods used in species distribution models (Liu *et al.* 2009). We also recommend those works of Pearce *et al.* (2000), Allouche *et al.* (2006), and Lobo *et al.* (2008) for comparative assessments and critical reviews on the subject.

## 15.6. Ensemble

Different levels of uncertainty exist in species distribution modelling techniques depending on the results of the assessment. For this reason, ‘consensus models’ based on consensus among modelling methods have been recently introduced (Thuiller, 2004). While Marmion *et al.* (2009) introduce several “consensus” methods, Thuiller *et al.* (2009) explain ensemble forecasting of species distributions using the ‘biomod’ R-package. Modela uses the ‘biomod2’ R-package (evolution of ‘biomod’ R-package). The following are the available ensemble types:

- ‘PA\_dataset+repet’
- ‘PA\_dataset+algo’
- ‘PA\_dataset’
- ‘algo’
- ‘all’

where ‘PA\_dataset’ indicates that pseudo-absence data sets will be used as ensemble factors. Note that it is possible to create several pseudo-absence data sets (PA) of the same size but from different locations. This is because the methods for pseudo-absence data selection are random within an area or the whole work frame.

‘repet’ indicates that a model repetition will be used as an ensemble factor. Modela allows obtaining repetitions from the same model (algorithm) that can be evaluated according to different criteria. To perform the evaluation, the data set should be divided. Each division generates new sets of presence and pseudo-absence data with different responses (repetitions using the same model). For this reason, ‘repet’ can be used as an ‘ensemble factor’.

‘algo’ indicates that the algorithm(s) selected will be used as an ensemble factor for modelling.

Finally, “all” indicates that all the above, “PA\_dataset”, “repet” and “algo”, will be used together to assemble models. Modela allows the user to assemble either all models generated or only those models that exceed a minimum evaluation value.

We recommend reading the document entitled 'Ensemble Modelling: the different available ways to build ensemble forecasts' for a better insight into the theory and implications of each ensemble forecast option. This document can be downloaded from the following site <https://www.capfitogen.net/es/EnsembleModelingAssembly.pdf>.

## 15.7. Predictions under future projected bioclimatic data

To quantify the potential impact of climate change on plant species, it has been proposed to project models into GIS layers representing future climate scenarios obtained through GCMs such as those in Table 2. Thanks to the model, a pattern is established between a species' presence/pseudo-absence and ecogeographical variables (bioclimatic variables corresponding to current conditions). The pattern is then projected into layers showing future projected bioclimatic variables. If any current soil and/or geophysical layers have been used to establish the model, both future (for bioclimatic) and current (edaphic and/or geophysical) layers are to be included in the future projection. Such future projections have allowed predicting the impact of climate change on CWR of peanut (*Arachis*), potato (*Solanum*), and cowpea (*Vigna*) in Latin America and Africa (Jarvis *et al.*, 2008). Studies of this kind require comparisons between the present and future projections, and these are possible in Mcompare.

## 15.8. FIGS subsets through modelling (calibration technique)

Chapter 11 describes the FIGS technique to identify germplasm (subsets) highly likely to have traits of interest for breeders. Using the 'calibration technique', Modela tool can identify FIGS subsets. The tool carries out a modelling process assuming presence-absence data. The presence data set (table) corresponds to accessions that have been assessed for the desired trait; the absence data set (table) will include those accessions that do not have the trait of interest. The pattern obtained in the modelling is projected into a different set of accessions not evaluated for the desired trait. Modela also projects into GIS layers to obtain a map showing the probability of the trait to appear.

## 15.9. How does Modela work?

To understand the complexity behind Modela and how it works, Fig. 52 illustrates the relationships and processes involved to obtain a predictive map of species distribution. On the other hand, Fig. 53 illustrates the processes and elements involved to obtain FIGS subsets through the calibration method.

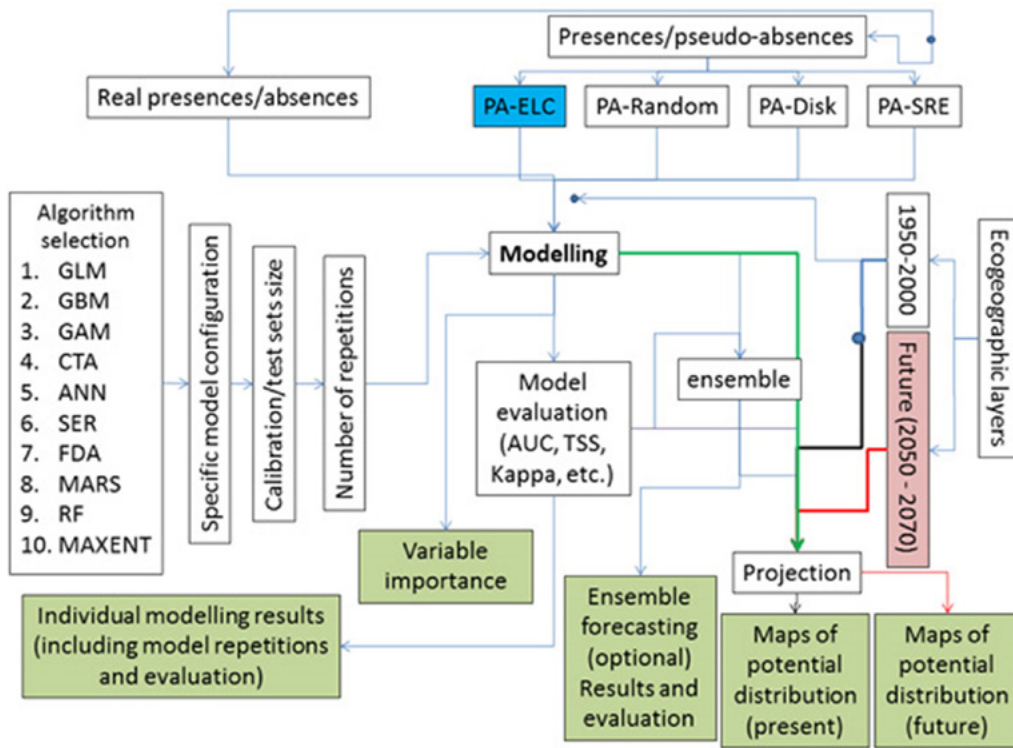


Figure 52. Diagram illustrating Modela's process of generating models and species distribution projections.

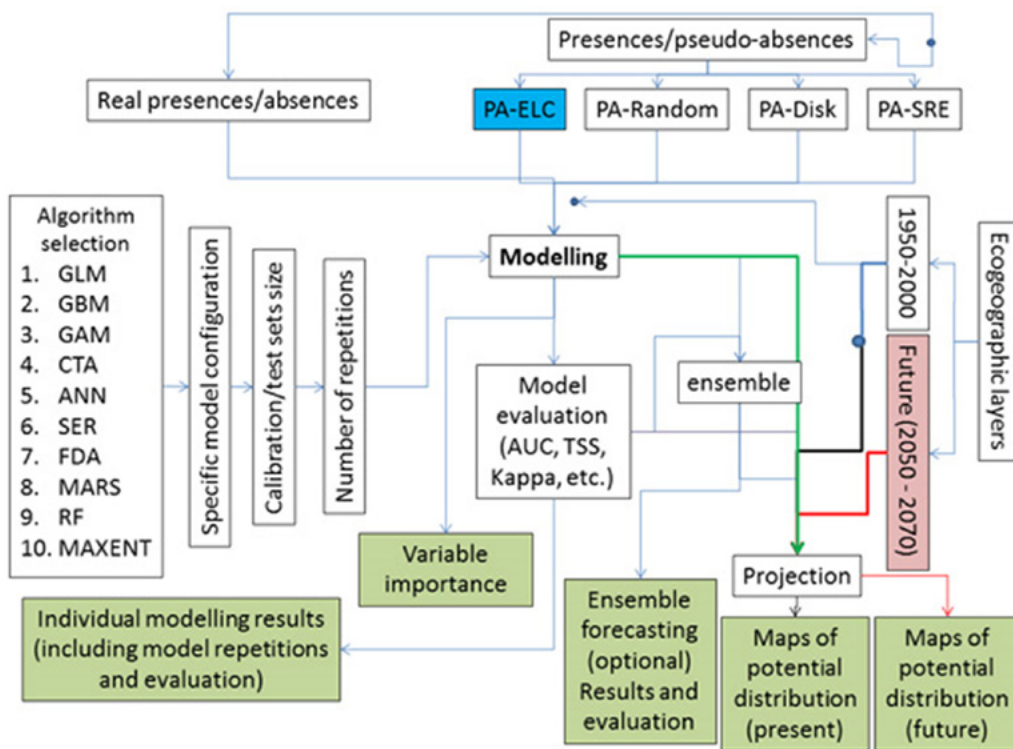


Figure 53. Diagram illustrating Modela's process of identifying FIGS subcollections.

### 15.9.1. Running Modela and SelecVIF using a table of preset parameters

It is important to highlight a new feature available in Modela and SelecVIF that is not included in any of the other tools. Here, the user can select the options in each parameter by using a template provided and without the time limit imposed by the virtual server. Through this option (enabled for both, *local mode* and *on server mode*) Modela and SelecVIF can be set up before the tools' interface is opened. It will require an Excel table ('TableParameters.xlsx' file for Modela or 'TableParametersSelecVIF.xlsx' for SelecVIF) that can be downloaded from [https://drive.google.com/drive/folders/16Fl9V9STi2iZ4vUHDkO\\_rhbXLk-7i99e?usp=sharing](https://drive.google.com/drive/folders/16Fl9V9STi2iZ4vUHDkO_rhbXLk-7i99e?usp=sharing) or that can be found inside the "ModelOptions" folder (path CAPFITOGEN3/ModelOptions). These Excel files contain a spreadsheet named "ParameterTable" that includes all the parameters that control the use of Modela.

In this table, the user specifies the value (column 'Valor/Value') for each parameter of Modela or SelecVIF. The cells of column 'Valor/Value' are set up to give the users only the possible and valid options for each parameter. Fig. 54 shows the table 'ParameterTable' within the Excel file 'TableParameters.xlsx'. An example is provided on how the user can fill in the different fields in column 'Value (indicate here your option)'. The different colours in column 'Parameter' indicate the relationship between the different parameters. Thus, fields with the same colour have a certain degree of dependence or subordination. Column 'Is this parameter required?' indicates to the user (as each parameter is being set up) whether he/she should provide information for the following parameters or not (in descending order). Fig. 54 shows how the user determines the TRUE value in parameter 'geoqual'. At that moment, in column C ('Is this parameter required?'), the message in front of parameter 'totalqual' changes from 'totalqual value is not necessary' to 'totalqual value must be provided'. In column 'Notes', the type of data to be inserted is described for each parameter. Some information that may be helpful to the user is also added in this column.

Parameter	Value (indicate here your option)	Is this parameter required?	Notes
1 ruta	C:/CAPFITOGEN3	Ok	para indicar la ruta use sólo este slash/   to in
2 pais	Argentina	Ok	Selecione un país/región / Select one country
3 pasaporte	Pasaporte.txt	Ok	nombre incluye la extensión .txt, el archivo de
4 geoqual	TRUE	Ok	TRUE = Si   yes, FALSE = No
5 totalqual		90 totalqual value must be provided	número entre 0 y 100   number from 0 to 100
6 distdup		1 Ok	número (por defecto 0)   number (0 by default)
7 temp	present	Ok	Selecione una opción   Select one option
8 gcm	present	gcm (CIMP6) cell must be filled	Selecione un modelo global climático de futuro
9 rcp/ssp	future	rcp/ssp value must be provided	Selecione un escenario SSP de emisiones   Se
10 proy		proy value must be provided	Selecione un periodo de proyección para datos
11 resol1		resol1 value is required	Selecione una opción   Select one option
12 variabv[1]		Please select one variable for variabv[1]	Selecione una variable ecogeográfica a ser utilizada
13 variabv[2]		Please select one variable for variabv[2]	Selecione una variable ecogeográfica (no seleccionada)
14 variabv[3]		Please select one variable for variabv[3]	Selecione una variable ecogeográfica (no seleccionada)
15 variabv[4]		Select additional variable for variabv[4] if necessary	Selecione una variable ecogeográfica (no seleccionada)
16 variabv[5]		Select additional variable for variabv[5] if necessary	Selecione una variable ecogeográfica (no seleccionada)
17 variabv[6]		Select additional variable for variabv[6] if necessary	Selecione una variable ecogeográfica (no seleccionada)
18 variabv[7]		Select additional variable for variabv[7] if necessary	Selecione una variable ecogeográfica (no seleccionada)
19 variabv[8]		Select additional variable for variabv[8] if necessary	Selecione una variable ecogeográfica (no seleccionada)
20 variabv[9]		Select additional variable for variabv[9] if necessary	Selecione una variable ecogeográfica (no seleccionada)
21 variabv[10]		Select additional variable for variabv[10] if necessary	Selecione una variable ecogeográfica (no seleccionada)
22 variabv[11]		Select additional variable for variabv[11] if necessary	Selecione una variable ecogeográfica (no seleccionada)
23 variabv[12]		Select additional variable for variabv[12] if necessary	Selecione una variable ecogeográfica (no seleccionada)
24 variabv[13]		Select additional variable for variabv[13] if necessary	Selecione una variable ecogeográfica (no seleccionada)

Figure 54. Spreadsheet (Parameter table) included in the Excel file "TableParameters.xlsx".

Some multi-value parameters, such as 'variabv', which is partially shown in Fig. 54, repeat themselves several times so that the user can choose different options. For instance, 20 repetitions (numbered as follows: [1], [2], [3], etc.) are available for 'variabv' meaning that the user can select up to 20 ecogeographic variables for modelling. The fact that several repetitions are available does not mean that options have to be selected for all of them. If only 3 variables are needed for modelling, only the first three variables (variabv[1], variabv[2], and variabv[3]) will be selected, leaving the rest blank. In the same Excel file ('TableParameters.xlsx') the user will find a second spreadsheet named 'Variables\_Codes'. The table in this spreadsheet will help the user to connect each 'variabv' code to its description in both English and Spanish.

All required parameters must be filled according to the case and type of analysis to be performed, indicating the options up to the last parameter ('resultados') for *local mode*. In *on server mode*, the definition of parameters 'ruta' and 'resultados' in the Excel format does not generate any action since the results will always be saved in the User's Files and Results area. Finally, when the Excel format is complete, it is recommended to save it in the same format (.xlsx). The table 'ParameterTable' must be saved in tab-delimited text format (.txt) for the tool to be able to read it. In *local mode*, the text file with the parameter table must be saved in the 'ModelOptions' folder. Parameter 'partablename' will appear in both modes; In *on server mode*, the user must upload the table in tab-delimited text format for 'partablename', whereas, in *local mode*, the parameter script must indicate the name of the file.

### **15.9.2. Algorithm set up**

The algorithms used in the R-package biomod2 (parameter 'modelby', option 'biomod') can be set up in detail to the requirements of the case or the user needs.

Follow these steps to set up the options/parameters of an algorithm:

- Open the 'ModelOptions\_OpcionesModelos.xlsx' file in the 'ModelOptions' folder (path CAPFITOGEN3/ModelOptions).
- This Excel file contains 10 spreadsheets, each of them with the name of an algorithm. On each spreadsheet, there is a table with different options/parameters to be specified for the particular algorithm. Column 'Option' shows the original name of the option/parameter to set up, and column 'Value (Default)' displays the default value for that option/parameter (value that Modela will use in case FALSE was indicated in parameter 'modelopc'). Modify the options/parameters in column 'Value (Default)' according to the specific needs or requirements; drop-down lists will be available in some fields here facilitating the selection/modification (Fig. 54). It is highly recommended to study in detail the consequences of changing options/parameters of the algorithm. Thus, modelling errors (due to the use of out-of-range values) or unexpected results will be avoided. All the possible changes available are those according to the function 'BIOMOD.Model.Options' of the R-package biomod2. Therefore, it is advisable to review all information related to the options/parameters of the algorithms of the manual of the R-package biomod2 (downloadable from here <https://goo.gl/ogj359>).

	A	B
1	Option	Value (default)
2	distribution	bernoulli
3	n. trees	bernoulli
4	interaction.depth	huberized
5	n.minobsinnode	multinomial
6	shrinkage	adaboost
7	bag.fraction	poisson
8	train.fraction	coxph
9	cv.folds	quantile
10	keep.data	parwise
11	verbose	1
12	perf.method	3
		FALSE
		FALSE
		CV

**Figure 55.** Table to modify the options/parameters of the GBM algorithm. The table is included in the Excel file “ModelOptions\_OpcionesModelos.xlsx”. For the option “distribution”, a drop-down list shows the different options available.

- Save as tab-delimited text file each option/parameter table for each algorithm in which a change has been made. The modified tables have to be saved individually in the folder ‘ModelOptions’ (path CAPFITOGEN3ModelOptions). Be careful not to replace text files already saved there and named with the algorithm code plus the number 1 (e.g., ‘GLM1.txt’). These files will be used when the following three conditions take place simultaneously: 1) the specific algorithm has been selected (parameter ‘modelos’); 2) parameter ‘modelopc’ has been indicated that the modified options/parameters are to be used for some of the selected algorithms, and 3) no changes will be made in some of the algorithms (just for these unmodified algorithms, the files named with the codes of each algorithm plus the number 1 will be used for ‘default’ options. For example: GLM1.txt, GBM1.txt, GAM1.txt, etc.).
- In parameter ‘modelopc’, indicate TRUE (or check the interface’s selection box) if specific options/parameters for at least one algorithm are going to be used for modelling.
- Indicate the algorithms whose options/parameters will be changed. Select TRUE (or check the interface’s selection box) in the required parameter of Modela: ‘GLM’, ‘GBM’, ‘GAM’, ‘CTA’, ‘ANN’, ‘SRE’, ‘FDA’, ‘MARS’, ‘RF’ or ‘MAXENT’.
- Indicate in parameters ‘GLMopt’, ‘GBMopt’, ‘GAMopt’, ‘CTAopt’, ‘ANNopt’, ‘SREopt’, ‘FDAopt’, ‘MARSopt’, ‘RFopt’ or ‘MAXENTopt’ the name of the text file with the modified options/parameters table.
- Fill in the rest of the fields according to the criteria of your analysis.



### **15.9.3. Model projection on future bioclimatic layers**

Bfuture (Chapter 14) provides future projections of bioclimatic information as GIS layers ready to be used in other tools. Modela can model current ecogeographic data and project the models on future bioclimatic layers, along with current soil and geophysical layers. For Modela to project on future bioclimatic layers, just indicate the value of 'future' in parameter 'temp'. Also, select the GCM (parameter 'gcm'), the RCP (parameter 'rcp'), and the period of the projection (parameter 'proy') of the projected future layers to use. To project on future layers, the following is required:

- To have the same current and future layers for the country/region/personalized work frame ('user1', 'user2' or 'user3'), and at the same resolution (cell size). The following is an example of how to verify that the same current and future layers are available: we chose 'Bolivia' as the target country, 5x5 km resolution, and future layers downloaded and adjusted in Bfuture for the GCM 'GFDL-ESM2G', rcp4.5 and projected to 2050. The user should find layers (i.e., files with the name of the variable and extensions .grd and .gri) for current conditions in the path CAPFITOGEN3rdatamaps/bolivia/5x5/, and for future conditions in CAPFITOGEN3/rdatamapsf/Bolivia/GFDL-ESM2G\_45\_50/5x5/
- Not to select any monthly average variable (Temp prom1, Temp prom 2,...Temp prom 12) in parameter 'variabv'.

Finally, when the value of 'future' is indicated in parameter 'temp', Modela will generate not only future predictions but also predictions for both, current and future bioclimatic scenarios.

### **15.9.4. Keeping the proportion of presence/absence data or presence/pseudo-absence data in both 'train' and 'test' sets**

The assessment of the models created in Modela is optional when using the package dismo for modelling (parameter 'modelby', option 'dismo'). However, when modelling with biomod2 (parameter 'modelby', option 'biomod'), the assessment is completed by default. Nevertheless, when assessing the model(s) it is necessary to determine the proportion of presence/absence data or presence/pseudo-absence data that will define the 'train set' (parameters 'datadiv2' and 'datadiv' when modelling with 'dismo' and 'biomod', respectively). If using the R-package biomod2 for modelling, the function 'BIOMOD\_Modeling' will create the 'test' and 'train' sets by randomly assigning presence and absence data (or presence and pseudo-absence data); note here that it will not consider the final proportion of presence and absence data (or presence and pseudo-absence data) assigned to each set. Thanks to parameter 'modprop', in Modela the user can choose between two options. Either a random selection of presence and absence data (or presence and pseudo-absence data) by biomod2 or to keep the user's original proportions of presence and absence data (or presence and pseudo-absence data) in the 'test' and 'train' sets. If choosing the second option (parameter 'modprop', option TRUE), Modela will randomly select the necessary presence and absence data (or pseudo-absence data) to keep the original proportions in the 'test' and 'train' sets.

### **15.9.5. Steps and requirements in Modela**

Here are described the steps and requirements for the different analyses that can be run in Modela.

#### **15.9.5.1 Modelling using presence-absence data**

The following entry data are required for the analysis:

- a)** Presence data table in CAPFITOGEN passport format (parameter 'pasaporte'). Optionally, GEOQUAL tool could have been previously used to assess the quality of the georeferencing.
- b)** Absence data table in CAPFITOGEN passport format (parameter 'pasaporteb').

And the steps to follow for this type of analysis are:

- c)** Indicate the options for the following parameters: 'ruta', 'pais', 'distdup', and 'resol1'. The resolution of the ecogeographic layers (parameter 'resol1') will be that of the maps of potential distribution of the models.
- d)** If projecting on future bioclimatic layers (parameter 'temp', option 'future'), specify parameters 'gcm', 'rcp', and 'proy'.
- e)** Select the ecogeographic variables/layers for the modelling (parameters 'variabv', 'latitud', and 'longitud').
- f)** Select 'absence' in parameter 'ausencia'.
- g)** Select the option 'biomod' in parameter 'modelby' (option 'dismo' only provides algorithms for presence-only data). Also provide an ID, preferably a short word, for the modelling with 'biomod' (parameter 'modelid').
- h)** Select the algorithm(s) to be used (parameter 'modelos'). If required, use the specific options/parameters for the algorithms (parameter 'modelopc') saved in the individual text files. In this case, select TRUE in the parameter with the code of each algorithm with specific options. Then, type the name of the table (in text format) that contains the specifications for the selected algorithm. For example, if specific options are to be used for a GLM algorithm, first select TRUE in parameter 'GLM', and then type the name of the table in text format into parameter 'GLMopt'.
- i)** Set the number of evaluation repetitions (parameter 'modrep'), the proportion of presence/absence data for both 'test' and 'train' sets (parameter 'datadiv'), and the predominance of absence data over presence data, or vice versa (parameter 'preval'; see biomod2 manual for more details). If interested in estimating the importance of the predictors, set the number of permutations as well. Then, select the different options from the following parameters: 'testparam': choose the evaluators to be used; 'reescal': choose to resize the outcomes of the different algorithms to a scale of 0-1000; 'modcomp': choose to obtain results using the whole data set, without dividing the data into 'test' and 'train', and 'modprop': choose to use proportions of presence/absence in both 'test' and 'train' sets.
- j)** Select whether to generate projections (i.e., to obtain a predictive map) for all the models or only for those that exceed a threshold for specific evaluators (parameter 'proysome'). If choosing the latter, indicate which of the evaluators previously selected in parameter 'testparam' acts as primary (parameter 'proytest1'), and its threshold value (parameter 'proytest1u'). Optionally, the user can filter the models a second time by using a secondary parameter (parameter 'proysome2', option TRUE). Again, indicate both the evaluator to use (parameter 'proytest2') and the threshold value (parameter 'proytest2u').
- k)** Additionally, the user could obtain a binary potential distribution map (parameter 'binar') for one of the models. Use here one of the evaluators previously selected in parameter 'testparam' (indicating this in parameter

‘binarmet’) and using the evaluator’s optimized value as a cut-off value. It is also possible to create binary maps for all the remaining model projection maps in DIVA-GIS (option ‘Reclass’ under the ‘Grid’ tab) as the optimized cut-off values for all evaluators are available to the user in one of the result tables.

- l) The user can choose to obtain a ‘mask’ map. This map shows areas highlighted where the prediction is uncertain. Uncertainty here happens when the values of the layers for the ecogeographical variables are out of the range used to calibrate the models (parameter ‘maskout’).
- m) If the user wants to assemble all or some of the generated models, select TRUE in parameter ‘ensamb’. Now, in parameter ‘mod2bens’ select either ‘all’ to assemble all the models or ‘best’ to assemble only the models previously selected to be projected (i.e., those models that exceeded a threshold for a specific evaluator, parameters ‘proysome’, ‘proytest1’, ‘proytest1u’, ‘proysome2’, ‘proytest2’, and ‘proytest2u’). It is necessary to specify how the models will be assembled (see section 15.5) considering the repetitions, the pseudo-absence sets, and the algorithms used (parameter ‘tipensam’). If ‘best’ was selected in parameter ‘mod2bens’, then define the models to be assembled. In this case, indicate which evaluators are going to be used; either those selected in parameter ‘testparam’ (parameter ‘ensamet’, option FALSE) or the evaluators and thresholds previously used to select the models to project (parameter ‘ensamet’, option TRUE). The user may choose to assess the assembled model (parameter ‘testparam2’, option TRUE) in which he/she should specify the evaluators to be used (parameter ‘testparam3’, multiple choice). Next, indicate with some of the following parameters the method for combining the models to assemble: ‘probmean’, ‘probcv’, ‘probc1’ (if selecting this method, indicate parameter ‘probalfa’ as well), ‘probmedian’, ‘probca’ or ‘probmw’ (if selecting this method, indicate parameter ‘probmwd’ as well).
- n) Type the path to save the results (parameter ‘resultados’).

### 15.9.5.2 Modelling using presence/pseudo-absence data

The following entry data are required for the analysis:

- a) Presence data table in CAPFITOGEN passport format (parameter ‘pasaporte’). Optionally, GEOQUAL tool could have been previously used to assess the quality of the georeferencing.

And the steps to follow for this type of analysis are:

- a) Indicate the options for the following parameters: ‘ruta’, ‘pais’, ‘distdup’ and ‘resol1’. The resolution of the ecogeographic layers (parameter ‘resol1’) will be that of the predictive maps of the models.
- b) If projecting on future bioclimatic layers (parameter ‘temp’, option ‘future’), specify parameters ‘gcm’, ‘rcp’, and ‘proy’.
- c) Select the ecogeographic variables/layers for modelling (parameters ‘variabv’, ‘latitud’, and ‘longitud’).
- d) Select ‘pseudo-absence’ in parameter ‘ausencia’.
- e) Select the number of pseudo-absence sets to obtain (parameter ‘pareps’). Take into account here that for each set, a modelling process will be run. Also select the size of each set (parameter ‘pansel’) and the strategy for generating pseudo-absence data (parameter ‘pastrat’).
- f) If the user chose the strategy ‘sre’ in parameter ‘pastrat’, indicate the quantil value in parameter ‘pasreq’. If the strategy ‘disk’ was chosen instead, indicate both the minimum (parameter ‘padiskmin’) and maximum (parameter ‘padiskmax’) distances (meters). Or, if the ‘elc’ strategy was selected, indicate the following: frequency

of quartile(s) in parameter 'paclc' (see section 15.2.1), the name of the ELC map (parameter 'mapaclc'), and the file (.txt) that contains the table with the statistics of the ELC map (parameter 'stateclc').

- g) In parameter 'modelby', the user can select either 'biomod' or 'dismo'. If selecting 'dismo', choose the algorithm for modelling via 'dismo' in parameter 'dismodel' (it allows only one algorithm to be chosen). Optionally, indicate if you want to obtain a model with all presence/pseudo-absence data available (parameter 'modcomp1') and to do modelling with the assessment option (parameter 'diseval', option TRUE). If the latter is selected, indicate the percentage of presence/pseudo-absence data that will make up the 'train' set (parameter 'datadiv2'). Finally, type the path to save the results (parameter 'resultados'). If selecting 'biomod' instead, follow the steps h) to n) in section 15.8.5.1.

### 15.9.5.3 FIGS using the trait's presence-absence data

The following entry data are required for the analysis:

- a) Data table in CAPFITOGEN passport format with the accessions with the trait of interest (parameter 'pasaporte'). Optionally, GEOQUAL tool could have been used previously to assess the quality of the georeferencing.
- b) Data table in CAPFITOGEN passport format with the accessions that do NOT have the trait of interest (parameter 'pasaporteb').
- c) Data table in CAPFITOGEN passport format with accessions not characterized or assessed for the trait of interest (parameter 'pasaportec').

To run the analysis, follow the steps in section 15.8.5.1 ('Modelling using presence-absence data'). As an additional requirement, select TRUE in parameter 'figs'.

## 15.10. Maxent's algorithm

Please remember that Maxent is only to be used with presence-only data (presence/pseudo-absence). **The Maxent software can be used through the 'biomod' package only from CAPFITOGEN3.** To use Maxent, R has to 'make contact' first with the Java software and provide the latter with all the necessary information. The results will be then brought back to R, and after doing the necessary programming, they will be saved into the folder indicated in parameter 'resultados', along with all the projections from the other algorithms used.

The option to make models with the dismo package from Modela is disabled from CAPFITOGEN3. Although some parameters of Modela (remnants of previous versions) still appear in the option 'dismo', this should not be selected as it would generate an execution error.

### 15.10.1. Requirements for Maxent

The biomod2 package can build Maxent models if they are run externally from their original Java program. To install the required Java software, follow the instructions below:

1. Install the following version of Java development kit (jdk) from: [https://drive.google.com/file/d/0B3ZXTUGKr-jX\\_b01BUkpINzVTQTg/view?usp=sharing](https://drive.google.com/file/d/0B3ZXTUGKr-jX_b01BUkpINzVTQTg/view?usp=sharing)

Double-click on the file jdk-8u60-windows-x64.exe that is generated when unzipping the downloaded .zip file. Follow the option 'next' in each step of the process until the installation is finished. This file will install a Java version for Windows 64 bit.

2. In Windows, go to Settings and search for 'environment variables'; select the option 'Edit the system environment variables'. This will open a pop-up window called 'System Properties'; find and click on the 'Environment variables' button and there create (or edit, if they are already created) the following variables in the 'User variables for pc' window:
  - JAVA\_HOME: for this variable, define the path C:\Program Files\Java\jre1.8.0\_60 or the equivalent path that takes you to the Java Run Environment (or JRE) program you have just installed.
  - JRE\_HOME: define the same path for JAVA\_HOME.
  - path: define the path to R 64 bit; it can be C:\Program Files\R\R-3.6.3\bin\x64
3. In the 'System variables' section, there is a variable called 'Path'. Click on it and then on the 'Edit ...' button. A window called 'Edit system variable' will appear, with a list of paths in it. Create a new path by clicking on the 'New' button, paste here the path that leads to the Java SE Development Kit (or JDK) bin folder, which is usually C:\ProgramFiles\Java\jdk1.8.0\_60\bin (including the 'bin' part). Once this new path has been created, it must be positioned as the first one in the list, locating it there with the help of the 'Move Up' button.
4. Type the word java in Windows 'Run' (program in Start-All Programs-Accessories). A window will appear with a text box preceded by a prompt that says 'Open:'. After typing 'java' in the box and clicking 'OK', a black background window will open and close quickly and no error window should appear. If so, the environment variables set in step 2 are correct.

## 15.11. Using Modela tool

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and Modela tool has been selected, a series of parameters must be specified by the user. Modela and SelecVIF represent an exception in the way the parameters are entered. The number of parameters that must be set up for these tools is high, and for many of the parameters, the options are limited and depend on the configuration of other parameters. For this reason, the user has Excel files with tables that facilitate the configuration of the parameters. All the parameters in Modela are listed below, including those that are shared with SelecVIF (the name of the parameter appears accompanied by - SelecVIF). A few parameters that are unique to SelecVar are listed afterward.

### 15.11.1 Initial parameters defined by the user

#### 15.11.1.1 Parameter: ruta (only for local mode) - SelecVIF

Explanation: Path where CAPFITOGEN3 tools have been copied or are to be found. Avoid spaces in this path (for

example 'C:/Mis documentos'). Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/ or C:/CAPFITOGEN3 or D:/MisHerramientas/CAPFITOGEN3, etc.

#### 15.11.1.2 Parameter: pais - SelecVIF

Explanation: Select the country/region or customized area (options user 1, user 2, or user 3 produced by rLayer and Bfuture tools) where all or most of the presence data occur. If presence data occur in more than one country, you may select a region, subcontinent, continent, or customized areas rather than an individual country.

#### 15.11.1.3 Parameter: pasaporte - SelecVIF

Explanation: Type the name of the file containing the occurrence table (in CAPFITOGEN passport tab-delimited format). Do not forget to include the file extension (.txt). For example, if the file is named 'table', you should write 'table.txt'. Please remember that this file must first be saved in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. In case you want to perform a FIGS analysis, the table for parameter 'pasaporte' corresponds to that of accessions positively evaluated for the trait.

#### 15.11.1.4 Parameter: geoqual - SelecVIF

Explanation: Select this option (indicating TRUE) if the presence/passport data table has been analysed using GEOQUAL tool and, thus, contains 5 additional columns (SUITQUAL, LOCALQUAL, COORQUAL, TOTALQUAL, and TOTALQUAL100). Please use the table produced by GEOQUAL, usually called PasaporteOriginalEvaluadoGEOQUAL.txt, as a passport table in the point above.

#### 15.11.1.5 Parameter: totalqual - SelecVIF

Explanation: If your passport table was obtained from GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers from 0 (zero quality) to 100 (maximum quality).

#### 15.11.1.6 Parameter: distdup - SelecVIF

Explanation: Determine the distance (in km) under which you consider that two collection sites represent the same population (geographical duplicate). A value of zero (the minimum value by default) excludes from the analysis occurrences with identical coordinates. The greater the distance, the higher the number of occurrences that would be considered geographical duplicates.

#### 15.11.1.7 Parameter: temp

Explanation: Select one option here if the modelling will be performed using present (1950-2000 period) bioclimatic data (please select 'present') or present and future (projected to 2050 or 2070, please select 'future'). If all the

variables (predictors) you wish to use to model correspond to edaphic or geophysical variables, select the option 'present'. Note that the option 'future' requires present and future bioclimatic data available for the country, region, or customized area selected in parameter 'pais'. Future bioclimatic data must be located in the 'rdatamapsf' folder where Bfuture usually locates its results.

#### 15.11.1.8 Parameter: rcp/ssp

Explanation: Applies only if 'temp' = 'future'. Select the representative concentration pathway for the projected bioclimatic data you have available for the country/region (parameter 'pais') in the 'rdatamapsf2' folder. RCP is, in other words, the four greenhouse gas concentration trajectories adopted by the IPCC: RCP2.6 (here appears as '26'), RCP4.5 (45), RCP6 (60) and RCP8.5 (85). These options are related to radiative values in the year 2100 compare to pre-industrial era values (+2.6, +4.5, +6.0, and +8.5 W/m<sup>2</sup>, respectively). There are four available SSP: 126 (ssp1, 2.6), 245 (ssp2, 4.5), 370 (ssp3, 7.0) and 585 (ssp5, 8.5).

#### 15.11.1.9 Parameter: gcm

Explanation: Applies only if 'temp' = 'future'. Select the global climate model (GCM or general circulation model) for the projected bioclimatic data you have available for the country/region (parameter 'pais') in the 'rdatamapsf' folder. These models simulate the future weather scenarios according to the assumed atmospheric concentration of greenhouse gases (rcp/ssp). If you want to know more about GCM, please visit <http://goo.gl/4XhU6g> or <http://goo.gl/2VUuRP>.

#### 15.11.1.10 Parameter: proy

Explanation: Applies only if 'temp' = 'future'. Select the period for the bioclimatic projected data you have available for the country/region (parameter 'pais') in the 'rdatamapsf2' folder. Option 50 represents the value/year 2050 (average for 2041-2060) and option 70 represents the value/year 2070 (average for 2061-2080).

#### 15.11.1.11 Parameter: resol1 - SelecVIF

Explanation: Select the resolution level you wish to use for the variables (predictors) to be modelled and the resultant maps. Note that 1x1 km offers greater resolution but requires greater computing capacity than 5x5 km; however, this is not as limiting a factor as it is for ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents, continents, or customized areas.

#### 15.11.1.12 Parameter: variabv - SelecVIF

Explanation: Select the ecogeographical variables to be used as model predictors. If you wish to model distributions for present and future bioclimatic conditions (option 'future' in parameter 'temp') these projected bioclimatic variables for future conditions have to be available in the 'rdatamapsf' folder for the working area selected (country, region, or customized area selected in parameter 'pais'). IMPORTANT NOTE: If you wish to project the models for future bioclimatic conditions (option 'future' in parameter 'temp') please DO NOT select here mean monthly temper-

atures (from 'Temp prom 1' to 'Temp prom 12') since these variables are not available in WorldClim and, therefore, in the 'rdatamapsf' folder.

#### **15.11.1.13 Parameter: *latitud* - SelecVIF**

Explanation: Select this option (indicating TRUE) if you wish to include latitude as a predictor in your model(s).

#### **15.11.1.14 Parameter: *longitud* - SelecVIF**

Explanation: Select this option (indicating TRUE) if you wish to include longitude as a predictor in your model(s).

#### **15.11.1.15 Parameter: *genero* - SelecVIF**

Explanation: Type the name of the genus of the species for which you want to model the distribution. This field is mandatory, but the term entered here does not need to match the field GENUS content in the CAPFITOGEN passport table. The term entered here will only be used to name (together with parameter 'species') some result folders.

#### **15.11.1.16 Parameter: *especie* - SelecVIF**

Explanation: Type the name of the epithet of the species whose distribution you want to model. This field is mandatory, but the term entered here does not need to match exactly the content of field SPECIES in the CAPFITOGEN passport table. The term entered here will be used only to name (along with parameter 'genero') some result folders.

#### **15.11.1.17 Parameter: *ausencia* - SelecVIF**

Explanation: Select the type of absences you wish to use for modelling. Option 'absence' means you will use true absence data to model, providing tables with absence data (parameter 'pasaporteb'). If you want to perform a FIGS analysis to select germplasm, you must also select option 'absence'. If you do not have true absence data for modelling, please select option 'pseudo absence'. Note 1: If you have selected option 'dismo' for parameter 'modelby', then parameter 'ausencia' will automatically be 'pseudoabsence'. Note 2: If you have selected option 'absence' in parameter 'ausencia', option 'biomod' for parameter 'modelby', and option 'MAXENT' in parameter 'modelos', this last algorithm (Maxent) will not be used for modelling.

#### **15.11.1.18 Parameter: *pasaporteb* - SelecVIF**

Explanation: Applies only if 'ausencia' = 'absence'. Type the name of the file containing the absence data table (in CAPFITOGEN passport tab-delimited format) without forgetting to include the file extension (.txt). For example, if the file is named 'table', you should write 'table.txt'. Please remember that this file must first be saved in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. In case you want to perform a FIGS analysis, table for parameter 'pasaporteb' corresponds to accessions that do not possess the trait of interest.



### 15.11.1.19 Parameter: **figs**

Explanation: Applies only if 'ausencia' = 'absence'. Mark this option if you want to perform a FIGS analysis. To perform this analysis, you must provide a table with positively evaluated accessions (parameter 'pasaporte'), a table with negatively evaluated accessions (parameter 'pasaporteb'), and a table with non-evaluated accessions to project the models (parameter 'pasaportec').

### 15.11.1.20 Parameter: **pasaportec**

Explanation: Applies only when 'figs' has been checked. Type the name of the file containing the non-evaluated accessions (in CAPFITOGEN passport tab-delimited format) without forgetting to include the file extension (.txt). For example, if the file is named 'table', you should write 'table.txt'. Please remember that this file must first be saved in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. This table is only valid for FIGS analysis (parameter 'figs') and contains accessions to project the models.

### 15.11.1.21 Parameter: **pareps - SelecVIF**

Explanation: Applies only if 'ausencia' = 'pseudo absence'. Determine the number of pseudo-absence subsets you wish to generate for modelling. Note that selecting option 'dismo' for parameter 'modelby', only the first set of pseudo-absences will be used (in this case please type here 1). If you select 'modelby' = 'biomod', the number of modelling processes for each algorithm (parameter 'modelos') will be according to multiplication of 'pareps' by 'modrep'. Consider that a high number of modelling processes can take a long processing time.

### 15.11.1.22 Parameter: **pansel - SelecVIF**

Explanation: Applies only if 'ausencia' = 'pseudo absence'. Determine the number of pseudo-absences per set (number of sets determined in parameter 'pareps'). This number corresponds to the number of pseudo-absences to be used for modelling along with occurrence data (parameter 'pasaporte'). It is very important to consider a balance between presence and pseudo-absence data when modelling. For more information see <http://goo.gl/4aMsDZ>.

### 15.11.1.23 Parameter: **pastrat - SelecVIF**

Explanation: Applies only if 'ausencia' = 'pseudo absence'. Select the strategy to obtain pseudo-absences. Note that some strategies could generate smaller sets of pseudo-absences (parameter 'pansel') since some of those strategies reduce the area available to select the requested number of pseudo-absences.

Option 'random' Selects pseudo-absences from any 'background' cell (cells that cover the work frame and have ecological information but no presence data).

Option 'sre' means pseudo-absences will be selected from outside the broadly defined environmental conditions for the species, by first creating an environmental envelope model defined by a specific quantile ('pasreq' parameter).

Pseudo-absences will be selected outside this envelope.

Option 'disk' creates circular buffer areas (disks) around presence sites using minimum (parameter 'padiskmin') and maximum (parameter 'padiskmax') distances from each presence site. Pseudo-absences will be selected within the disk area.

Option 'elc' uses species frequency quartiles from categories of an ELC map (previously obtained using ELCmaps tool). The user can select which frequency quartiles (parameter 'paelc') will be used to extract pseudo-absences.

#### 15.11.1.24 Parameter: padiskmin - SelecVIF

Explanation: Applies only if 'pastrat' = 'disk'. Type here the distance (in meters) from each presence site to the inner circle of the disk that will be used to select pseudo-absences.

#### 15.11.1.25 Parameter: padiskmax - SelecVIF

Explanation: Applies only if 'pastrat' = 'disk'. Type here the distance (in meters) from each presence site to the outer circle of the disk that will be used to select pseudo-absences.

#### 15.11.1.26 Parameter: pasreq - SelecVIF

Explanation: Applies only if 'pastrat' = 'sre'. Type here the quantile value to define the size of the environmental envelope for the species and the outer space to select pseudo absences. This value must range from 0 to 0.5.

#### 15.11.1.27 Parameter: mapaelc - SelecVIF

Explanation: Applies only if 'pastrat' = 'elc'. Enter the name of the file containing the ELC map (previously obtained with ELCmapas tool) that should be in the ELCmapas folder, one of the folders that make up the CAPFITOGEN directory. This map must be in DIVA-GIS format (.grd extension, exactly as produced by the ELCmapas tool) and its name must include the file extension. Thus, if the name of the map is 'mapa\_elc\_spain', you should enter 'mapa\_elc\_spain.grd'.

#### 15.11.1.28 Parameter: statelc - SelecVIF

Explanation: Applies only if 'pastrat' = 'elc'. Enter the name of the file with the descriptive statistics table of the ELC map produced by the ELCmapas tool (the tool usually names this kind of file as 'Estadist\_ELC\_' plus the name of the country or region). Like the ELC map, this file should also be located in the ELCmapas folder. Similarly, the name should be followed by the file extension (.txt as the file is a table). Therefore, if the file is called 'Estadist\_ELC\_spain', you should enter 'Estadist\_ELC\_spain.txt'.

**15.11.1.29 Parameter: paelc**

Explanation: Applies only if 'pastrat' = 'elc'. Select one or several frequency quartiles to define areas to be used for random pseudo-absences generation.

**15.11.1.30 Parameter: modelby (disabled parameter or that should only indicate 'biomod')**

Explanation: 'biomod' is the only option that applies. 'biomod' uses biomod2 package (<https://goo.gl/ogj359>) for modelling. biomod2 in CAPFITOGEN offers nine algorithms for presence-absence or presence-only data (parameter 'modelos' = 'GLM','GBM','GAM','CTA','ANN','SRE','FDA','MARS','RF'), and one for presence-only data ('MAXENT') (requires 'ausencia' = 'pseudo absence'). All of them are customizable (parameter 'modelopc') and they can be assembled (parameter 'ensamb'). Option 'biomod' also allows you to use all sets of pseudo-absences to model (parameter 'pareps' can be =1 or >1).

**15.11.1.31 Parameter: dismodel**

Explanation: Applies only if 'modelby' = 'dismo'. Select the algorithm for modelling under the dismo package environment.

**15.11.1.32 Parameter: modcompl**

Explanation: Applies only if 'modelby' = 'dismo'. Mark this option if you wish to obtain an additional model using 100% of your presence data.

**15.11.1.33 Parameter: diseval**

Explanation: Applies only if 'modelby' = 'dismo'. Mark this option if you wish to obtain models by dividing the presence data into train (to model) and test (to evaluate models) sets.

**15.11.1.34 Parameter: datadiv2**

Explanation: Applies only if 'modelby' = 'dismo'. Type here a number between 1 and 100 to determine the percentage of presence data for obtaining the model (train set). The remaining presence data will be used for the test set.

**15.11.1.35 Parameter: modelid**

Explanation: Applies only if 'modelby' = 'biomod'. Type here a name to identify this modelling process in relation to other past or future processes.

#### 15.11.1.36 Parameter: modelos

Explanation: Applies only if 'modelby' = 'biomod'. Select one or more algorithms to be used for modelling. Model codes explanation can be found at <https://goo.gl/ogj359> or in the CAPFITOGEN user manual. Selecting 'MAXENT' requires the use of only-presence data. Also, a Java file must be downloaded and properly located from [https://biodiversityinformatics.amnh.org/open\\_source/maxent/](https://biodiversityinformatics.amnh.org/open_source/maxent/) (please see the user manual).

#### 15.11.1.37 Parameter: modelopc

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to model using specific parameters for the algorithms selected. Keep in mind that selecting this option, you must provide text (.txt) files that define the parameters for each algorithm. These files must be located in the 'ModelOptions' folder. In case you do NOT check this option, the modelling process will use default parameters for each algorithm selected (these default parameters are listed in the excel file 'ModelOptions\_OpcionesModelos.xlsx' located in the 'ModelOptions' folder).

#### 15.11.1.38 Parameter: GLM

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the GLM (Generalized Linear Model) algorithm.

#### 15.11.1.39 Parameter: GLMopt

Explanation: Applies only when 'GLM' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for GLM, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

#### 15.11.1.40 Parameter: GBM

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the GBM (Generalized Boosting Model) algorithm.

#### 15.11.1.41 Parameter: GBMopt

Explanation: Applies only when 'GBM' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for GBM, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

#### 15.11.1.42 Parameter: GAM

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the GAM (Generalized Additive Model) algorithm.

#### 15.11.1.43 Parameter: GAMopt

Explanation: Applies only when 'GAM' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for GAM, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

#### 15.11.1.44 Parameter: CTA

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the CTA (Classification Tree Analysis) algorithm.

#### 15.11.1.45 Parameter: CTAopt

Explanation: Applies only when 'CTA' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for CTA, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

#### 15.11.1.46 Parameter: ANN

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the ANN (Artificial Neural Network) algorithm.

#### 15.11.1.47 Parameter: ANNopt

Explanation: Applies only when 'ANN' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for ANN, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

#### 15.11.1.48 Parameter: SRE

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the SRE (Surface Range Envelop or usually called BIOCLIM) algorithm.

#### 15.11.1.49 Parameter: SREopt

Explanation: Applies only when 'SRE' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for SRE, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

**15.11.1.50 Parameter: FDA**

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the FDA (Flexible Discriminant Analysis) algorithm.

**15.11.1.51 Parameter: FDAopt**

Explanation: Applies only when 'FDA' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for FDA, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

**15.11.1.52 Parameter: MARS**

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the MARS (Multiple Adaptive Regression Splines) algorithm.

**15.11.1.53 Parameter: MARSopt**

Explanation: Applies only when 'MARS' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for MARS, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

**15.11.1.54 Parameter: RF**

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the RF (Random Forest) algorithm.

**15.11.1.55 Parameter: RFopt**

Explanation: Applies only when 'RF' has been marked. Type the name of the file (including the .txt extension) that contains the list of parameters for RF, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

**15.11.1.56 Parameter: MAXENT**

Explanation: Applies only when 'modelopc' has been marked. Check this option if you wish to specify parameters for the MAXENT (Maximum Entropy) algorithm.

**15.11.1.57 Parameter: MAXENTopt**

Explanation: Applies only when 'MAXENT' has been marked. Type the name of the file (including the .txt extension)

that contains the list of parameters for MAXENT, some of them duly modified by you. This text file can be obtained from the excel file 'ModelOptions\_OpcionesModelos.xlsx' and must be located in the 'ModelOptions' folder.

#### 15.11.1.58 Parameter: modrep

Explanation: Specify the number of repetitions of model evaluations (evaluation process will be 'modrep' times repeated). Please consider that a high number for parameters 'modrep' and 'pareps' and the simultaneous use of several algorithms ('modelos') involves performing several modelling processes. This takes considerable time and requires high computing capacity.

#### 15.11.1.59 Parameter: datadiv

Explanation: Applies only if 'modelby' = 'biomod'. Type here a number between 1 and 100 to determine the percentage of presence data for obtaining the model (train set). The remaining presence data will be used for the test set.

#### 15.11.1.60 Parameter: preval

Explanation: Applies only if 'modelby' = 'biomod'. Determine the prevalence value. It allows giving more or less weight to some particular observations. If 'preval' = NA, each observation (presence or absence) has the same weight regardless of the number of presences and absences (this is the default option). If 'preval' = 0.5, absences will be weighted equally to presences. If 'preval' is set below or above 0.5, absences or presences are given more weight, respectively.

#### 15.11.1.61 Parameter: imporvar

Explanation: Applies only if 'modelby' = 'biomod'. Specify the number of permutations to estimate variable (predictors) importance from the modelling process. If you type here 0 (zero), the importance of the variable will not be estimated.

#### 15.11.1.62 Parameter: testparam

Explanation: Applies only if 'modelby' = 'biomod'. Select one or several model evaluation methods you wish to apply.

#### 15.11.1.63 Parameter: reescal

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to rescale all modelling algorithm predictions using a binomial GLM (0 to 1000 scale, as required by the Mcompare tool). If you do not check this option, prediction will be in the original scale of each modelling algorithm.

#### 15.11.1.64 Parameter: modcomp

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to obtain an additional model using 100% of your presence data.

**15.11.1.65 Parameter: modprop**

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to keep the original balance in your presence/absence or presence/pseudo-absence data when creating sets to train the model (train set) or test it (test set). Otherwise, train and test sets will be shaped with presence/absence or presence/pseudo-absence data in which the original balance may or may not be kept. In any case the selection of presence/absence or presence/pseudo-absence data for the train and test sets is random.

**15.11.1.66 Parameter: proysome**

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to project only the models that have exceeded an evaluation threshold (both, the evaluation method and threshold will be defined by the user). If you do not check this option, all models will be projected.

**15.11.1.67 Parameter: proytest1**

Explanation: Applies only when 'proysome' has been marked. Select an evaluation method (as a primary filter) to filter models to be projected. For this first evaluation method, a threshold must be set (parameter 'proytest1u').

**15.11.1.68 Parameter: proyest1u**

Explanation: Applies only when 'proysome' has been marked. Type here the value for the threshold for the primary evaluation method. Only models that exceed this threshold will be projected. Note that the optimum performance value is usually 1.

**15.11.1.69 Parameter: proysome2**

Explanation: Applies only when 'proysome' has been marked. Check this option if you wish to filter again the set of models to be projected using another evaluation method and another associated threshold.

**15.11.1.70 Parameter: proytest2**

Explanation: Applies only when 'proysome2' has been marked. Select an evaluation method (as a secondary filter) to be used to filter again the models to be projected. For this second evaluation method, a second threshold must be set (parameter 'proytest2u').

**15.11.1.71 Parameter: proytest2u**

Explanation: Applies only when 'proysome2' has been marked. Type here the value of the threshold for the secondary evaluation method. Only models that exceed primary and secondary thresholds will be projected. Note that the optimum performance value is usually 1.



**15.11.1.72 Parameter: binar**

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to obtain a binarized response map (values of 0 for low or null probability and 1 for high presence probability) based on the binarization process used by the selected evaluation method (parameter 'testparam').

**15.11.1.73 Parameter: binarmet**

Explanation: Applies only when 'binar' has been marked. Select the evaluation method whose binarization process you wish to use to obtain a binarized response map. This method must be listed in the selected methods in parameter 'testparam'.

**15.11.1.74 Parameter: maskout**

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to obtain a mask (map) that will identify locations where predictions are uncertain because the values of the variables (predictors) are outside the range used for calibrating the models.

**15.11.1.75 Parameter: ensamb**

Explanation: Applies only if 'modelby' = 'biomod'. Mark this option if you wish to assemble the models produced. You can assemble either all models produced or those with the best evaluations.

**15.11.1.76 Parameter: mod2bens**

Explanation: Applies only when 'ensamb' has been marked. Please indicate if you wish to assemble all models produced ('all') or the best-evaluated models ('best'). It corresponds to parameter 'chosen.models' of function BIOMOD\_EnsembleModeling in biomod2 package.

**15.11.1.77 Parameter: tipensam**

Explanation: Applies only when 'ensamb' has been marked. Select the way models will be assembled. To understand the implications of each option, please check the document 'EnsembleModelingAssembly.pdf' that can be downloaded from <https://www.capfitogen.net/es/EnsembleModelingAssembly.pdf> (the same document is included in the 'Documentación\_Referencias/Modela' folder). It corresponds to parameter 'em.by' of function BIOMOD\_EnsembleModeling in biomod2 package.

**15.11.1.78 Parameter: ensamet**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use the same set of evaluators and thresholds that were used to select the models to be projected (parameters 'proytest1' and 'proytest1u',

optionally 'proytest2' and 'proytest2u') to select the models to be assembled. If this option is not checked, all available evaluators will be used. The evaluators and their thresholds used in this option will be assigned to parameters 'eval.metric' and 'eval.metric.quality.threshold' (respectively) of function BIOMOD\_EnsembleModeling in biomod2 package.

#### **15.11.1.79 Parameter: testparam2**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to evaluate the assembled models.

#### **15.11.1.80 Parameter: testparam3**

Explanation: Applies only when 'ensamb' has been marked. Select one or several model evaluation methods you wish to apply. It corresponds to parameter 'models.eval.meth' of function BIOMOD\_EnsembleModeling in biomod2 package.

#### **15.11.1.81 Parameter: probmean**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use the mean probabilities across predictions (models) as ensemble method.

#### **15.11.1.82 Parameter: probcv**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use the coefficient of variation across predictions (models) as ensemble method.

#### **15.11.1.83 Parameter: probci**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use the confidence interval around the mean probabilities across predictions (models) as ensemble method.

#### **15.11.1.84 Parameter: probalfa**

Explanation: Applies only when 'probci' has been marked. Please indicate the significance level for estimating the confidence interval.

#### **15.11.1.85 Parameter: probmedian**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use the median of probabilities across predictions (models) as ensemble method.

**15.11.1.86 Parameter: probca**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use a committee averaging as an ensemble method. For this method, all models are binarized. The committee averaging score is the average of binary predictions (like a simple vote). Each model votes for the species presence or absence. For each site, the sum of 1 is then divided by the number of models. In this way, this method offers both a prediction and uncertainty measure. When the prediction is close to 0 or 1, it means that most of the models agree to predict the absence or presence of the species, respectively. When the prediction value is around 0.5, it means that half the models predict the presence of the species and the other half its absence.

**15.11.1.87 Parameter: probmw**

Explanation: Applies only when 'ensamb' has been marked. Check this option if you wish to use the weighted sum of probabilities across predictions (models) as an ensemble method.

**15.11.1.88 Parameter: probmwd**

Explanation: Applies only when 'probmw' has been marked. This parameter defines the relative importance of the weights. Option 'proportional' will assign proportional weights to the evaluation scores that each model obtained in the modelling process.

**15.11.1.89 Parameter: resultados (only for local mode) - SelecVIF**

Explanation: Insert the path to the folder where the results of the analysis will be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

**15.11.2 SelecVIF exclusive parameters****15.11.2.1 Parameter: testNAs**

Explanation: Indicate if you want (TRUE) or not (FALSE) to remove the presences or collection sites that do not extract information before the modelling process.

**15.11.2.2 Parameter: cutoff**

Explanation: Enter the value that indicates the VIF threshold. Variables above this value will be discarded.

### 15.11.2.3 Parameter: dynamic

Explanation: Indicate whether you want (TRUE) or not (FALSE) to filter variables by VIF through a step-by-step process, eliminating high VIF variables in each step and determining again the VIF of the remaining variables.

## 15.12. Results of Modela

Once all the parameters have been defined, click on the ‘Analyse’ button. The whole processes of pseudo-absence generation (if required), modelling, model projection and ensemble (if required) could take from 1 minute to hours. Large work frames, high resolutions (small cell size), and the use of a high number of algorithms and evaluation methods can slow down the process and/or place great demands on system resources.

Describing in detail all the results generated by Modela tool would be too complex. Therefore, the information below shows how to organize the results, how to make up the names of the folders, and the types of tables and maps that can be obtained.

The results folder includes the text file ‘Parametros.Parameters.Modela.txt’ containing the different options for the parameters used in modelling, and a set of different folders in which all the results from Modela are organized. These different folders are the result of the different types of presence/absence data (presence/pseudo-absence or presence/absence) that can be observed in Table 5, and the different modelling options (Table 6).

**Table 5.** Folders created in the path specified in parameter ‘resultados’. The different folders are the result of the different types of presence/absence data used. Columns ‘ausencia’ and ‘pastrat’ correspond to the parameters specified in sections 15.11.1.19 and 15.11.1.23. The names in red change according to the specifications in parameters ‘genero’ and ‘especie’.

Options “ausencia”	Options “figs” or “pastrat”	Folders	Code
absence	figs=FALSE	Real_Absences_genero_especie	1
absence	figs=TRUE	Absences_NonEval_FIGS_genero_especie	2
pseudo absence	sre	PseudoAbsences_SRE_genero_especie	3
	disk	PseudoAbsences_Disk_genero_especie	4
	elc	PseudoAbsences_ELC_genero_especie	5
	random	PseudoAbsences_genero_especie	6

**Table 6.** Folders created in the path specified in parameter ‘resultados’. The folders here are the result of modelling with *dismo* (parameter *modelby* = ‘*biomod*’). The names in red change according to the specifications in parameters ‘*genero*’, ‘*especie*’, and ‘*modelid*’.

Options	Folders	Code
Basic	<i>genero.especie</i>	10
	Results_Evaluation_Models_ <i>modelid</i>	11
	Results_Project_raw_ <i>modelid</i>	12
imporvar > 0 (section 15.11.1.63)	Results_Variables_importance_ <i>modelid</i>	13
maskout = TRUE (section 15.11.1.76)	Results_Clamping_mask_ <i>modelid</i>	14
binar = TRUE (section 15.11.1.74)	Results_Project_binar_ <i>modelid</i>	15
ensamb = TRUE (section 15.11.1.77)	Results_Ensemb_Individual_Models_ <i>modelid</i>	16
ensamb = TRUE y mod2bens = best (sections 15.11.1.77 y 15.11.1.78)	Results_Ensemb_Best_Model_ <i>modelid</i>	17
ensamb = TRUE y testparam2= TRUE (sections 15.11.1.77 y 15.11.1.81)	Results_Evaluation_Ensemble_ <i>modelid</i>	18
temp = future (section 15.11.1.9)	Results_Project_raw_ <i>modelid</i> _F	19
temp = future and maskout = TRUE, ensamb = TRUE, mod2bens = best, binar = TRUE	Results_Clamping_mask_ <i>modelid</i> _F	20
	Results_Project_binar_ <i>modelid</i> _F	21
	Results_Ensemb_Individual_Models_ <i>modelid</i> _F	22
	Results_Ensemb_Best_Model_ <i>modelid</i> _F	23
figs = TRUE and ensamb = TRUE (sections 15.11.1.21 and 15.11.1.77)	Results_Emsemb_non_evaluated_germplasm	24

In the above tables, it is possible to associate each folder with its content easily as the folders are named according to the parameters they are based on. From now on, and to make things easier, when referring to a folder we will mention the code on the right side of each table.

The folder ‘*genero.especie*’ (originating from parameters ‘*genero*’ and ‘*especie*’) is created only when modelling with *biomod2* (parameter *modelby* = ‘*biomod*’). It is important to mention here that inside this folder there is a set of folders and files generated by ‘*biomod2*’ R-package that are useful for technicians familiarized with the package. However, for the purpose here, any detailed explanation about the actual content seems unnecessary as most of these results can only be visualized in R.

### 15.12.1. Modela tables

#### 15.12.1.1. Presence/absence or presence/pseudo-absence data tables

Most of the tables in Modela tool results can be found in folders 1 to 6. These folders correspond to presence and absence or pseudo-absence data used for modelling. Table 7 lists all the possible tables that can be found in folders 1 to 6.

**Table 7.** Files that correspond with the type of data (presence/absence or presence/pseudo-absence data) used for modelling.

Options	File name	Folder code	Table description
Any modelling option	input_data.txt	1,2,3,4,5,6	Table showing both, presence and absence/pseudo-absence data used for modelling, and the ecogeographical variables
Any modelling option	extraction_table_PresAbs.txt extraction_table_PresAbs.xls extraction_table_Random.txt extraction_table_Random.xls extraction_table_Sre.txt extraction_table_Sre.xls extraction_table_Disk.txt extraction_table_Disk.xls extraction_table_ELC.txt extraction_table_ELC.xls	1,2,3,4,5,6	Table showing the values extracted for each ecogeographic layer for presence and absence/pseudo-absence data (indicated in column 'inputs.data.species' as 1 and 0 or NA, respectively)
ausencia='pseudo absence' and pastrat='elc'	Genbank_ELC.txt Genbank_ELC.xls	5	Table showing all fields of the passport table plus an additional column (BGcat) with the ELC map category extracted for each site
ausencia='pseudo absence' and pastrat='elc'	PATable.txt PATable.xls	5	Table showing information to identify the generated pseudo-absence data (column resvar, values = NA) that are part of each set of repetitions (columns PA1, PA2, PA3,... PAN), TRUE=included, FALSE=excluded
ausencia='pseudo absence' and pastrat='elc'	Quartile_ELCmap_Classification.txt Quartile_ELCmap_Classification.xls	5	Table showing the reclassified values, percentages, frequencies, and classes (quartiles) that allow the selection of areas for obtaining pseudo-absence data in the ELC maps

### 15.12.1.2. Evaluation data tables

If the modelling has been performed with biomod2 (parameter modelby = 'biomod'), only the table 'models\_evaluation\_results.txt' (or 'models\_evaluation\_results.xls') will be obtained in both folders 11 and 18 if the ensemble option was chosen. This table contains the following columns:

- Model.name: assigns a compound name for each model assessed. In parameter ausencia = 'pseudo-absence' the name will be genero.especie\_PAX\_RUNY\_ZZZ, where X is the pseudo-absence data set, Y the repetition of the model, and ZZZ the algorithm used. If the modelling has been performed with presence-absence (parameter ausencia = 'absence'), the element RUNY is replaced by AllData.
- Eval.metric: specifies the evaluation method.
- Testing.data: value obtained for the evaluation method by the model.
- Cutoff: optimized threshold value determined for the evaluation method.
- Sensitivity: sensitivity parameter.
- Specificity: specificity parameter.

### 15.12.1.3. FIGS predictive data table

The table for the prediction of the trait's occurrence is generated in folder 24 under the file name 'ensembled\_prediction\_nonevaluated\_germplasm.txt' (or extension .xls for Excel). This table shows the coordinates of the accessions for which the models have been projected (columns 'DECLONGITUDE' and 'DECLATITUDE'). Here, the user will also find columns with the prediction values for each original or assembled model; these columns are named using the elements specified in section 15.12.2.2. Prediction values range from 0 to 1000 (values close to 1000 indicate a high probability of the trait to occur).

## 15.12.2. Modela maps

In practical terms, predictive maps are the most important results that can be obtained with Modela tool. However, Modela can also generate different kinds of maps to help analyse the results. These are maps in either raster (.tif or .grd extensions) and/or vector (shapefiles) format and can be found in some of the folders specified in Tables 5 and 6. All maps generated/created by Modela tool can be visualized and/or modified in DIVA-GIS software (<http://www.diva-gis.org>).

### 15.12.2.1. Presence/absence or presence/pseudo-absence maps

Table 8 shows details of the presence/absence or presence/pseudo-absence maps obtained.

**Table 8.** Files that correspond to maps showing where the presence/absence or presence/pseudo-absence data are located. These maps also show (if required) the areas where pseudo-absences were obtained via `pastrat= 'elc'`.

Options	File name	Folder code	Map description
According to the options selected in parameters 'ausencia' and 'pastrat'	Final_extraction_PresAbs.shp Final_extraction_Random.shp Final_extraction_Sre.shp Final_extraction_Disk.shp Final_extraction_ELC.shp	1,3,4,5,6	Point vector map showing presence and absence/pseudo-absence data. These are indicated in column 'inputs_dat' as 1 and 0, respectively. It also shows the extracted values for each site from the ecogeographic layers
ausencia= 'pseudo absence' and pastrat='elc'	Presences_Pseudoabsences_ELC.shp	5	Point vector map showing presence and pseudo-absence data. Columns PA1, PA2, PA3...to PAn (where n: 'True'/'False' value of parameter 'pareps') show the distribution of the sites in each pseudo-absence data set
ausencia= 'pseudo absence' and pastrat='elc'	mapa_Class_Sp.grd	5	Raster map showing the classification of the ELC map in quartiles of frequency
ausencia= 'pseudo absence' and pastrat='elc'	PA_availab.grd	5	Raster map showing the available areas (value 1) for obtaining pseudo-absence data according to the specifications in parameter 'paelc'
ausencia= 'absence' and figs=TRUE	points_to_be_predicted.shp	2	Vector map showing locations of non-assessed accessions

#### 15.12.2.2. Understanding the names built for the maps

The file names of the different maps (i.e., predictive maps, binarized or not, projected on present or future bioclimatic layers) are built combining specific elements that indicate what each map represents.

Below, the origin and meaning of the elements that make up the names of raster maps are described:

- **AllData:** Only found in folders 12, 15, 16, 19, 21, or 22 when the option 'absence' has been indicated in parameter 'ausencia'. It indicates that the map corresponds to a model that used presence/absence data.
- **PA:** Often found in folders 12, 15, 19 and 21. It indicates that pseudo-absences were used. If followed by a number (e.g., PA1), it indicates that the map corresponds to a model that used one of the pseudo-absence data



sets (the number indicates the set).

- **complete or full:** Often found in folders 7, 8, 9, 12, or 19. It indicates that the map corresponds to a model obtained with all available presence/absence or presence/pseudo-absence data.
- **F or Fut:** Only found in folders 7.2, 8.2, 9.2, 19, 20, 21, 22, or 23, and when indicating the option 'future' in parameter temp. It indicates that the map corresponds to a projection on future bioclimatic layers.
- **binar or binary:** It can be found in folders 7.2, 8.2, 9.2, 15, or 21. It indicates that the map corresponds to a binary projection using a threshold. If found together with the name of an evaluation method, it means that the threshold used comes from the actual evaluation method. For example, 'binaryROC' indicates that ROC's optimized threshold was used to create the binary projection.
- **RUN:** Only found in folders 12 and 19, and when biomod2 is used for modelling. It indicates that the map corresponds to a projection using one of the model's repetitions (set in parameter "modrep").
- **EMxxByXXX:** Only found in folder 16 when models have been assembled (parameter ensamb=TRUE); and also in folder 22 if parameter temp='future'. It indicates the function that has been used to assemble the models. After 'EM', the user can find any of the following options: mean, cv, cilnf and ciSup, median, ca, and wmean. These options refer to the ensemble function determined in parameters: probmean, provcv, probci, probmedian, probca, and probmw, respectively. After 'By', one of the following options will be observed: KAPPA, TSS, ROC, FAR, SR, ACCURACY, BIAS, POD, CSI, and ETS. The options refer to the evaluation method used.
- **Clamping\_mask\_map:** Found in folders 14 and 20. It is the mask map that identifies areas of uncertain prediction, and is generated by parameter maskout=TRUE
- **GLM, GBM, GAM, CTA, ANN, SRE, FDA, MARS, RF, or MAXENT:** Indicate the algorithm used for modelling. They are found when biomod2 is used for modelling. They usually constitute the last part of the names of the map files in folders 12 and 19 (projection maps scale 0-1000), and the middle part of the name of the map files in folders 15 and 21 (binary projection maps).
- **mergedXXXX:** Only found in folder 16 when models have been assembled (parameter ensamb=TRUE), and also in folder 22 if parameter temp='future'. Indicates the strategy chosen (parameter 'tipensam') to combine the models and build the assembled ones. After 'merged' you can find the code of the method used for assembling (here shown as XXXX). The code can be: 'Run' (models assembled using the different modelling repetitions, parameter 'modrep'), 'Algo' (models assembled using the different algorithms used, parameter 'modelos') or 'Data' (models assembled using the different pseudo-absence data sets or PA\_dataset, parameter 'pareps'). More than one 'mergedXXXX' can be found together in the name of the map file indicating that a combination of methods has been used for assembling (options 'all', 'PA\_dataset+repet', 'PA\_dataset+algo' in parameter 'tipensam'). For example, if 'all' is selected, the file name of the assembled maps will be 'mergedAlgo\_mergedRun\_mergedData'.
- **Best:** Only found in folders 17 and 23 when 'best' is indicated in parameter mod2bens. It corresponds to the maps of the assembled models that exceed a threshold value for one of the evaluation methods used.

### **15.12.3. Deleting files and temporary folders**

This applies only to *local mode*. Temporary files are created when generating models and maps of different sizes (MB or GB). The files, which are automatically saved in the folder 'Users' (name of the folder in Windows) under a long path like C:\Users\pc\AppData\Local\Temp, can quickly fill the hard drive (usually unit C:\). Windows stores temporary files from different programs (software) in the folder 'Temp'. These files are also here arranged in folders with the name of the actual program. Once all processes in Modela or any other tool are completed, we recommend closing all applications related to CAPFITOGEN and go to the folder 'Temp'. Then, look for the R-software folder (often named 'R\_raster\_pc') and check its size. If the size of the folder is enough to slow down your PC, consider now deleting some of its contents, but not the actual folder. Deleting temporary files will not affect the normal functioning of the tools. However, make sure that when doing so all running processes in Modela, other tools, or applications in R are completed and closed.

Finally, when using the Maxent algorithm for modelling using biomod2 (parameters modelby='biomod' and modelos='MAXENT'), an empty folder is generated in the path indicated in parameter 'ruta' (section 15.11.1.3). This empty folder, named combining both the genus (parameter 'genero', section 15.11.1.17) and species names (parameter 'especies', section 15.11.1.18), can be deleted without any problem.

## **15.13. References**

Akinwande, M. O., Dikko, H. G., Samson, A. 2015. Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. *Open Journal of Statistics* 5(07): 754-767.

Allouche, O., Tsor, A., Kadmon, R. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 5(7): 1223-1232.

Barbet-Massin, M., Jiguet, F., Albert, C. H., Thuiller, W. 2012. Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution* 3(2): 327-338.

Barnes, L. R., Grunfest, E. C., Hayden, M. H., Schultz, D. M., Benight, C. 2007. False alarms and close calls: A conceptual model of warning accuracy. *Weather and Forecasting* 22(5): 1140-1147.

Barnes, L. R., Schultz, D. M., Grunfest, E. C., Hayden, M. H., Benight, C. C. 2009. Corrigendum: False alarm rate or false alarm ratio? *Weather and Forecasting* 24(5): 1452-1454.

Booth, T. H., Nix, H. A., Busby, J. R., Hutchinson, M. F. 2014. BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MAXENT studies. *Diversity and Distributions* 20(1): 1-9.

Bradter, U., Kunin, W. E., Altringham, J. D., Thom, T. J., Benton, T. G. 2013. Identifying appropriate spatial scales of

predictors in species distribution models with the random forest algorithm. *Methods in Ecology and Evolution* 4(2): 167-174.

Chefaoui, R. M., Lobo, J. M. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecological modelling* 210(4): 478-486.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20: 37-40.

Elith, J., Leathwick, J. R., Hastie, T. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77(4): 802-813.

Elith, J., Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40(1): 677.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., Yates, C. J. 2011. A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17(1): 43-57.

Fawcett, T. 2004. ROC graphs: Notes and practical considerations for researchers. *Machine learning* 31: 1-38.

Fielding, A. H., Bell, J. F. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation* 24(01): 38-49.

Guisan, A., Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. *Ecological modelling* 135(2): 147-186.

Guisan, A., Edwards, T. C., Hastie, T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling* 157(2): 89-100.

Guisan, A., Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. *Ecology letters* 8(9): 993-1009.

Hijmans, R., Elith, J., 2015. Species distribution modeling with R. <https://rspatial.org/raster/sdm/>

Jarvis, A.; Williams, K.; Williams, D.; Guarino, L.; Caballero, P.J. Mottram, G. 2005. Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. *Genetic Resources and Crop Evolution* 52: 671-682.

Jarvis, A., Lane, A., Hijmans, R. J. 2008. The effect of climate change on crop wild relatives. *Agriculture, ecosystems & environment* 126(1): 13-23.

Liu, C., White, M., Newell, G. 2009. Measuring the accuracy of species distribution models: a review. In Proceedings 18th World IMACs/MODSIM Congress. Cairns, Australia (pp. 4241-4247).

Lobo, J. M., Jiménez-Valverde, A., Real, R. 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography* 17(2): 145-151.

Maiorano, L., Cheddadi, R., Zimmermann, N. E., Pellissier, L., Petitpierre, B., Pottier, J., Guisan, A. 2013. Building the niche through time: using 13,000 years of data to predict the effects of climate change on three tree species in Europe. *Global Ecology and Biogeography* 22(3): 302-317.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R. K., Thuiller, W. 2009. Evaluation of consensus methods in predictive species distribution modelling. *Diversity and distributions* 15(1): 59-69.

Mateo, R. G., Croat, T. B., Felicísimo, A. M., Munoz, J. 2010. Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions* 16(1): 84-94.

Özesmi, S. L., Özesmi, U. 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological modelling* 116(1): 15-31.

Parra-Quijano, M., Iriondo, J.M., Frese, L., Torres, E. 2012a. Spatial and ecogeographic approaches for selecting genetic reserves in Europe. In: Maxted, N., Dulloo, M.E., Ford-Lloyd, B.V., Frese, L., Iriondo, J., Pinheiro de Carvalho, M.A.A. (eds.) *Agrobiodiversity Conservation: securing the diversity of crop wild relatives and landraces*. CABI, Wallingford, UK

Parra-Quijano, M., Iriondo, J. M., Torres, E. 2012b. Applications of ecogeography and geographic information systems in conservation and utilization of plant genetic resources. *Spanish Journal of Agricultural Research* 2: 419-429.

Parra-Quijano, M., Iriondo, J. M., Torres, E. 2012c. Improving representativeness of genebank collections through species distribution models, gap analysis and ecogeographical maps. *Biodiversity and Conservation* 21(1): 79-96.

Pearce, J., Ferrier, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modelling* 133(3): 225-245.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., Ferrier, S. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications* 19(1): 181-197.

Ramirez-Villegas, J., Khoury, C., Jarvis, A., Debouck, D., Guarino, L. 2010. A gap analysis methodology for collecting crop genepools: a case study with Phaseolus beans. *PLoS One* 5(10): e13497. doi:10.1371/journal.pone.0013497.

Ramirez-Villegas, J., Jarvis, A., Läderach, P. 2013. Empirical approaches for assessing impacts of climate change on agriculture: The EcoCrop model and a case study with grain sorghum. *Agricultural and Forest Meteorology* 170: 67-78.

Roebber, P. J. 2009. Visualizing multiple measures of forecast quality. *Weather and Forecasting* 24(2): 601-608.

Russell, J., van Zonneveld, M., Dawson, I. K., Booth, A., Waugh, R., Steffenson, B. 2014. Genetic diversity and ecological niche modelling of wild barley: refugia, large-scale post-lgm range expansion and limited mid-future climate threats. *PLoS One* 9(2): e86021.

Scheldeman, X., van Zonneveld, M. 2011. *Manual de Capacitación en Análisis Espacial de Diversidad y Distribución de Plantas*. Bioersity International, Roma, Italia.

Sillero, N., Barbosa, A. M. 2021. Common mistakes in ecological niche models. *International Journal of Geographical Information Science* 35(2):213-226.

Thuiller, W., Araújo, M. B., Lavorel, S. 2003. Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* 14(5): 669-680.

Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. *Global Change Biology* 10(12): 2020-2027.

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M. B. 2009. BIOMOD—a platform for ensemble forecasting of species distributions. *Ecography* 32(3): 369-373.

VanDerWal, J., Shoo, L. P., Graham, C., Williams, S. E. 2009. Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological modelling* 220(4): 589-594.

Wood, J. M. 2007. Understanding and Computing Cohen's Kappa: A Tutorial. *WebPsychEmpiricist*. URL: [Journal at http://wpe.info/](http://wpe.info/).



Regional Workshop, Mexico (online training), March 2021.



# 16 | Mcompare Tool

## 16.1. Comparing current-future predictions

Scheldeman and van Zonneveld (2010) proposed an interesting comparison between predictions of species distributions using current and future climate predictors within a climate change scenario. For this comparison, current and future climate predictors (GIS layers) came from WorldClim (<http://www.worldclim.org>).

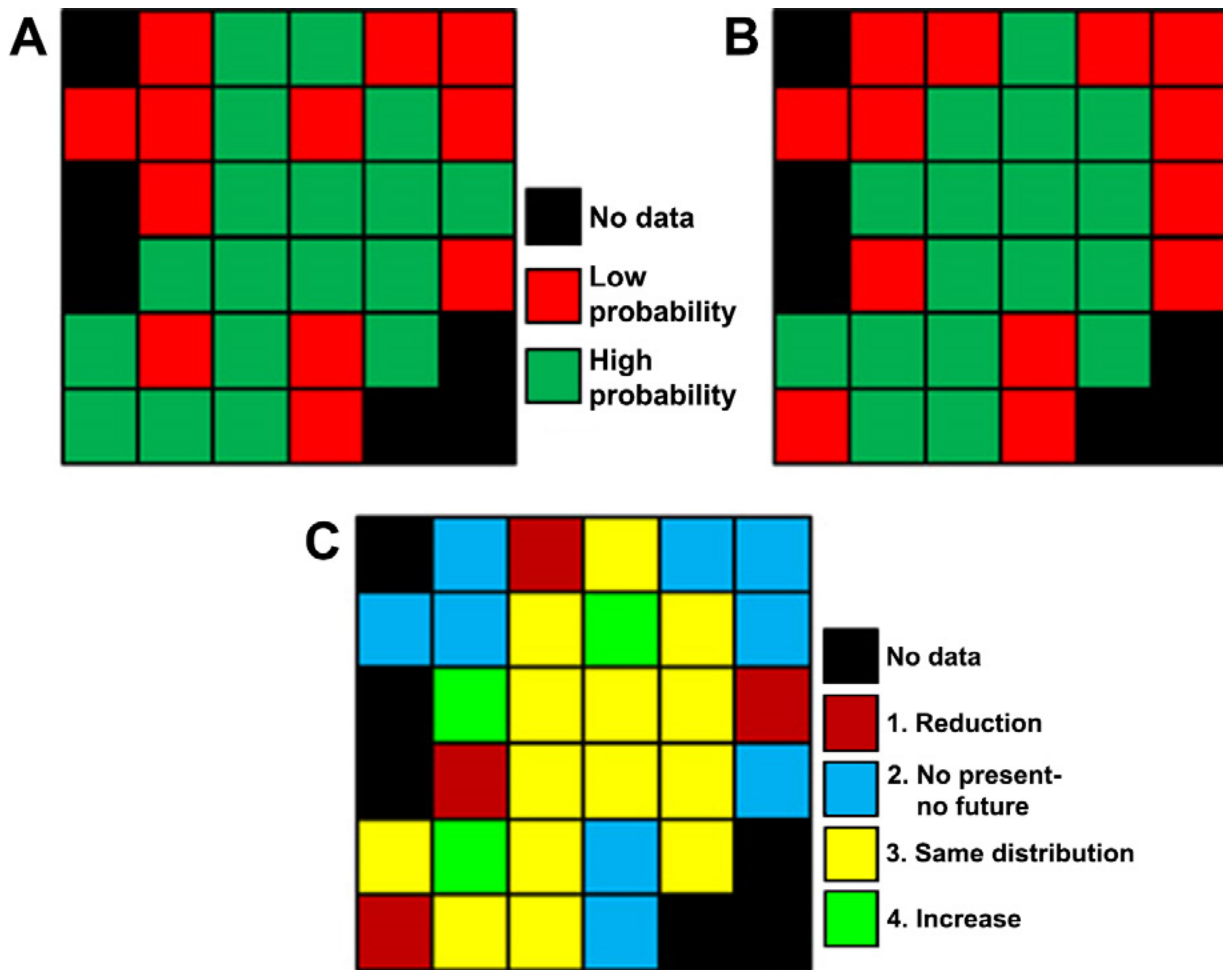
To obtain current and future projections, current data is modelled to obtain a pattern. Then, this pattern is projected on both GIS layers containing present data and GIS layers with future (projected) bioclimatic layers. If some soil or geophysical variables were selected as important predictors, the pattern would also be projected on soil and geophysical layers for current conditions (since it is considered that they are not going to change over relatively short periods of time). This process is carried out by the Modela tool (parameter 'temp', option 'future'). The GIS layers used by Modela to project potential distributions are the same as those used in the rest of CAPFITOGEN tools. The bioclimatic layers correspond to the WorldClim variables showing current data, and the GIS layers with future information are those downloaded and adapted by Bfuture, also from WorldClim.

Once the projections have been obtained, a comparison for each cell is carried out. Here, the aim is to detect the occurrence of any of the four situations outlined below by Scheldeman and van Zonneveld (2010) in their 'Training Manual on Spatial Analysis of Plant Diversity and Distribution':

1. Reduction of the distribution area: or 'High impact areas' by Scheldeman and van Zonneveld. Areas where a species potentially occurs in the present climate, but that will not be suitable in the future. They represent a threat to the survival of the species within the climate change scenario used.
2. No present - no future: or 'Areas outside of the realized niche' by Scheldeman and van Zonneveld. Areas (cells) that are not suitable for the species neither under current conditions nor under future (modelled) conditions. Within the climate change scenario used, they represent areas where the species occurs neither in this moment nor in the future.
3. Same distribution: or 'Low impact areas' by Scheldeman and van Zonneveld. Areas (cells) where the species can potentially occur in both present and future climates. They represent those areas where the species will still occur in the future regardless of climate change. This might be possible thanks to the species plasticity and/or adaptive potential.
4. Increase of the distribution area: or 'New suitable areas' by Scheldeman and van Zonneveld. Areas (cells) where a species could potentially occur in the future, but that are not suitable for natural occurrence under current conditions. They represent areas where the species does not currently occur, but it could in the future (expansion area). This is possible because the climate change scenario is beneficial or compatible with the species' adaptive potential in these areas.

Based on the comparison of current and future potential distributions, Fig. 56 shows how the above situations are classified cell by cell.

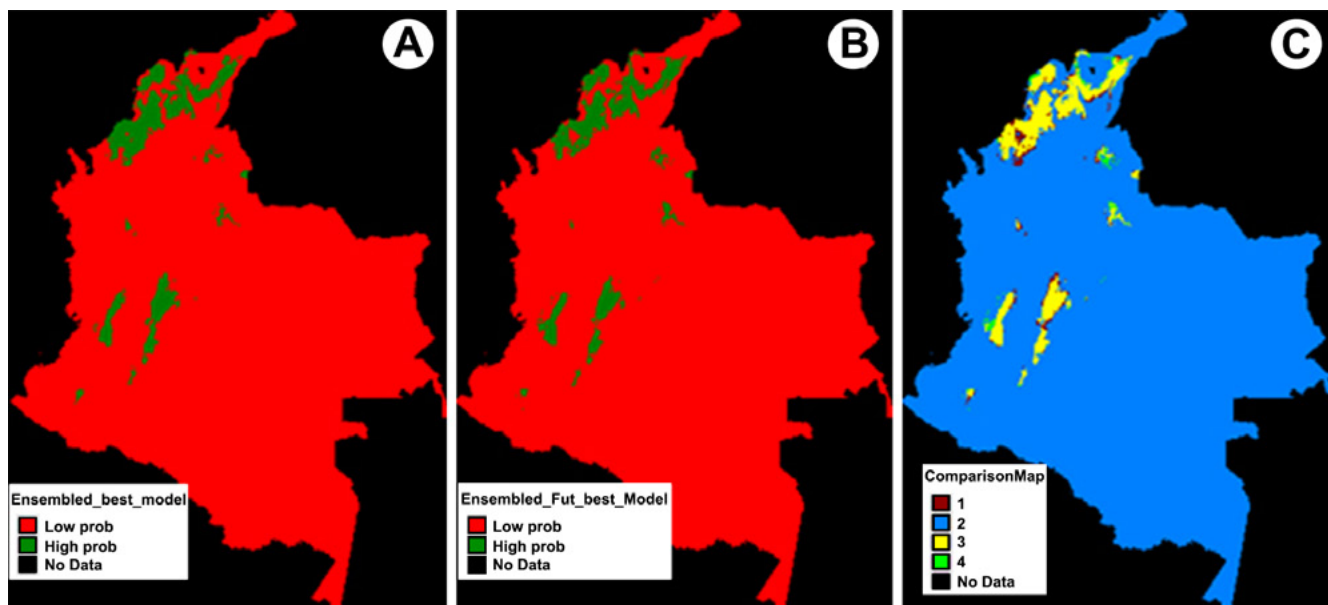




**Figure 56.** Comparison of two predictive layers, A) current conditions and B) future conditions. Situations described above (section 16.1) can be determined to obtain a layer C) in which each cell is classified according to the situation.

## 16.2. How does Mcompare work?

The user must indicate Mcompare where the maps are located and the names of the files. If the predictive maps (current or future) are not binary maps (see Chapter 15 for all related to binary maps) this must be indicated; then, the user must establish a cut-off value between 0-1000 to create binary maps. It is possible to use here the cut-off values provided in the model assessment tables (such as AUC, Kappa, or TSS). Additionally, the user must indicate the name of the file with the presence data in order to create a data point vector map (shapefile). This shapefile includes column 'Code Impact' as a result of the extraction from the situations' map (1 = reduction; 2 = no present-no future; 3 = same, and 4 = increase). of the cell value for each occurrence site.



**Figure 57.** Maps resulting from Mcompare. The colours used are the same as those in Fig. 56. Map of A) current potential distribution areas, B) future potential distribution areas, and C) classification of the four situations.

To do the comparison, the extension (xmin, xmax, ymin, and ymax) and cell size (1x1, 5x5, 10x10 or 20x20 km) of both maps of current and future potential distribution areas must match perfectly. The smallest difference between extensions or cell resolutions will automatically generate an error when using the tool.

## 16.3. Using Mcompare tool

Once CAPFITOGEN3 *local mode* tools are installed or CAPFITOGEN3 *on server mode* has been accessed and Mcompare tool is selected, a series of parameters must be specified by the user.

### 16.3.1 Initial parameters defined by the user

#### 16.3.1.1 Parameter: ruta (only for local mode)

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN3, C:/CAPFITOGEN3, D:/MisHeramientas/CAPFITOGEN3, etc

#### 16.3.1.2 Parameter: pasaporte

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add

the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

### 16.3.1.3 Parameter: geoqual

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if the passport data have been analysed using GEOQUAL tool and, thus, contain 50/51 columns (rather than the 45/46 columns in the basic passport model). To select this option, please use the table generated by GEOQUAL named 'PasaporteOriginalEvaluadoGEOQUAL.txt' (with this or any other name as long as it corresponds to this table) as a passport table (parameter 'pasaporte'). Selecting this option implies that accession data will be filtered, preserving the highest quality collection/occurrence sites in terms of their georeferencing.

### 16.3.1.4 Parameter: totalqual

Explanation: Applies only if 'geoqual' has been set as TRUE (✓ in *on server mode*). If your passport table has been previously analysed by GEOQUAL and you wish to consider a minimum quality standard to be met by data to be included in the analysis, determine the value of TOTALQUAL100 to be used as a threshold. The range covers values from 0 (zero quality) to 100 (maximum quality).

### 16.3.1.5 Parameter: disdup

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites represent the same population. The value zero (by default) excludes accessions with identical coordinates from the representativeness analysis. The determination of the distance depends on biological (gene flow) and spatial (mean population sizes) conditions. This is a specific parameter for the target species, and it will often be necessary to consult an expert for his/her concept.

### 16.3.1.6 Parameter: rutapresent (only for local mode)

Explanation: Type here the path to the raster map that contains the species distribution model for the target species using current bioclimatic data. Please do not use paths with spaces (for example C:/Mis documentos/). Note: use / (slash) instead of \ (backslash). For example, F:/, C:/CAPFITOGEN3, or D:/MisHerramientas/CAPFITOGEN3, etc.

### 16.3.1.7 Parameter: modelpresent

Explanation: Type here the name of the raster map that contains the species distribution model for the target species using current bioclimatic data. Do not forget to include the extension in the name. For example, 'Present\_complete\_

Species\_name.tif'. For *on server mode*, select the name of the raster map and upload it to the server by clicking on the button in front of the parameter.

#### 16.3.1.8 Parameter: **binarizedp**

Explanation: Indicate (TRUE in *local mode* or ✓ in *on server mode*) if the prediction on the present distribution model map is binarized (0 for absence, 1 for presence).

#### 16.3.1.9 Parameter: **binarthresp**

Explanation: Applies only if the option in 'binarizedp' has been set as FALSE (☐ in *on server mode*). Type a number from 1 to 999 to be used as a threshold to binarize the map for present predicted distribution. The map to be binarized should have prediction values from 0 to 1000. Values above the threshold indicated here will be assumed as presence and values below the threshold as absence.

#### 16.3.1.10 Parameter: **rutaigual (only for local mode)**

Explanation: Indicate with TRUE if the path for the model obtained using current bioclimatic data is the same for the model obtained using future bioclimatic data.

#### 16.3.1.11 Parameter: **rutafuture (only for local mode)**

Explanation: Applies only if the option in 'rutafuture' has been set as FALSE. If paths for present and future models are different, type here the path for the model obtained using future bioclimatic data. Please do not use paths with spaces (for example C:/Mis documentos/). Note: use / (slash) instead of \ (backslash). For example, F:/, C:/CAPFITOGEN3, or D:/MisHerramientas/CAPFITOGEN3, etc.

#### 16.3.1.12 Parameter: **modelfuture**

Explanation: Type here the name of the raster map that contains the species distribution model for the target species using future bioclimatic data (*local mode*). Do not forget to include the extension in the name. For example, 'Future\_complete\_species\_name.tif'. For *on server mode*, select the name of the raster map and upload it to the server by clicking on the button in front of the parameter.

#### 16.3.1.13 Parameter: **binarizedf**

Explanation: Indicate (TRUE in *local mode* or ✓ in *on server mode*) if the prediction on the future distribution model map is binarized (0 for absence, 1 for presence).

#### 16.3.1.14 Parameter: **binarthresf**

Explanation: Applies only if the option in 'binarized' has been set as FALSE (☐ in *on server mode*). Type a number from 1 to 999 to be used as a threshold to binarize the map for future predicted distribution. The map to be binarized should have prediction values from 0 to 1000. Values above the threshold indicated here will be assumed as presence and values below the threshold as absence.

#### 16.3.1.15 Parameter: **resultados**

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc.

## 16.4. Results of Mcompare

Once all the parameters and paths required by Mcompare are defined, the tool analysis process will start after clicking on the 'Run' button (*local mode* in RStudio) or 'Start' (*on server mode*).

After some time that may vary due to the type of analysis requested, Mcompare will save the results in the path and folder specified in 'resultados' (in *local mode*), or in the User's Files and Results area in *on server mode*.

The table 'Comparison\_stats.txt/xls' contains the four situations described in section 16.1. The number of cells for each situation in the map 'ComparisonMap.grd' is shown in column 'Freq'.

The table 'PresenceData\_Classification.txt/xls' contains a table built with the same column structure as that of the table 'Comparison\_stats.txt/xls'. However, the classification of the four situations does not correspond to the cells in the map 'ComparisonMap.tif' but to the presence sites represented in the map 'Current\_PresenceData\_Classified.shp'.

The raster map 'ComparisonMap.tif' (or 'ComparisonMap.grd' compatible with DIVA-GIS) shows the work frame classified according to the four possible situations described in section 16.1.

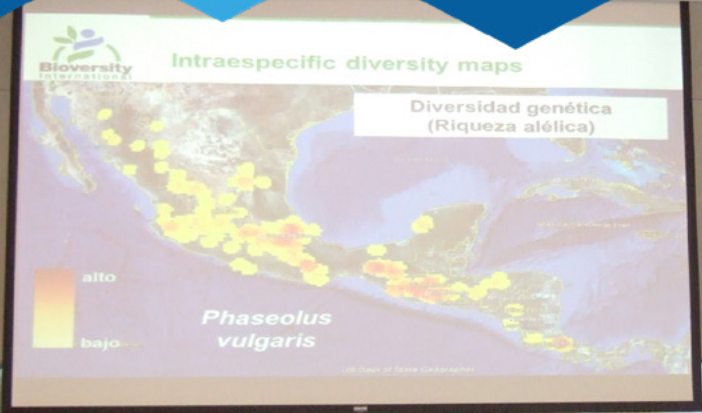
The point vector map 'Current\_PresenceData\_Classified.shp' shows the presence sites contained in the table previously indicated in parameter 'pasaporte'. Also, column 'Code Impact' is included in the shapefile's attribute table with the following codes for each situation: 1 = reduction, 2 = no present-no future, 3 = same, and 4 = increase.

Both types of maps can be visualized and edited (colours and values in the legend) in DIVA-GIS software.

## 16.5. References

Scheldeman, X., van Zonneveld, M. 2011. Manual de Capacitación en Análisis Espacial de Diversidad y Distribución de Plantas. Bioersity International, Roma, Italia.





1963  
50 años de normas  
internacionales  
para la conservación  
de la diversidad genética  
de las plantas  
cultivadas  
y sus recursos  
relacionados







# 17 | Tzones Tool

## 17.1. Conceptual basis

Species exhibit wide spatial and temporal variation in local patterns of genetic adaptation, which are related to several biotic and abiotic factors (Hanson *et al.*, 2017). The adaptation of a plant is regulated by the physiological and genetic variation of its populations and is determined by the environment to which they are exposed. Several anthropogenic factors influence climate dynamics causing changes in global temperature averages and altering precipitation patterns, with notable increases in extreme climate events and sea levels (Harris *et al.*, 2006; Williams and Dumroese, 2013). The asymmetric relationship between the increase in climate change associated phenomena and the adaptation capacity of plant species populations raises concerns about the risks to the *in situ* permanence of plant genetic resources. In this sense, the reproductive success of plant species will be limited by the alteration of environmental conditions, particularly in growing seasons (Chuine, 2010).

Faced with this challenge, ecological restoration practices can represent an opportunity to avoid the future loss of plant populations. The objective of ecological restoration is to re-establish an ecosystem that has been degraded, damaged, or destroyed by different factors, such as climate change. Proper or 'good' ecological restoration must include a broad approach that combines historical, social, cultural, political, aesthetic, and moral factors (Higgs, 1997). However, the technical issues that allow successful restorations must be figured out initially, since this is where genetic diversity among and within species plays a key role (Harris *et al.*, 2006).

As a result, the application of ecological restoration practices represents a challenge to protect plant biodiversity. Plants achieve adequate development and behaviour by growing under the specific environmental conditions in which they have evolved. One of the most common population restoration practices is the introduction of germplasm of the same species to reinforce the threatened population. Although the introduction of more adaptable individuals in affected populations would represent the introduction of foreign genetics that could displace or interbreed with local genetics, this technique may represent the only opportunity for populations in serious decline. These reinforcements will be successful if the genetic and adaptive compatibility is greater between the individuals to be reintroduced and the recipient population. Therefore, germplasm transfer requires strict control of genetic and environmental compatibility between source and destination populations. This has determined the development of approaches to carry out transfers that guarantee at least the compatibility of the environmental conditions of source and destination populations (adaptive profiles) and the potential effects of climate change (Potter & Hargrove, 2012).

## 17.2. Seed transfer zones

Seed Transfer Zones (STZ) are a methodological approach that has been developed to identify genetic resources of wild plants with appropriate adaptations to be used in the restoration of ecosystems affected by environmental changes (Havens *et al.*, 2015). Initially, STZs were based only on genetic compatibility between source and recipient populations to ensure effective reinforcement or restoration. Since genetic information on target wild plant species is often unavailable before collection or restoration activities, new STZ alternatives began to emerge. Thus, provisional STZs are proposed as a geographical area that includes two sites where the germplasm of one population can be transferred to another. A high environmental similarity between source and destination is sought, to minimize

the risk of maladaptation (Kramer & Havens 2009; Havens *et al.*, 2015). The term provisional STZ was proposed by Bower *et al.* (2014), who emphasize the need to provide temporary solutions to identify sources of germplasm for most cases in which there is no genetic information to guarantee source-receptor compatibility (particularly data on adaptive genetic variation). In addition to the lack of genetic information, another justification to provisional STZs is that abiotic adaptation information may indirectly reflect genetic variation (Peeters *et al.*, 1990).

The initial approaches to generate seed zones date from the beginning of the 20th century and had the purpose of avoiding or preventing failures in plant sowing -particularly forest species- (Bates 1930; Fowells 1949). Provisional STZs are based on assumptions about ecogeographic traits that are important for the species distribution range and are widely used by those running restoration programs (Omernik & Griffith 2014; Doherty *et al.*, 2017; Gibson & Nelson 2017; Germino *et al.*, 2019; Cevallos *et al.*, 2020). Some studies on genetic and phenotypic differentiation justify the use of provisional STZs and their bases for seed transfer (Johnson *et al.*, 2004; Doherty *et al.*, 2017; Durka *et al.*, 2017).

The background of provisional STZ is constituted by publications on germplasm transfer guidelines based on the use of ecoregion maps as a parameter for delineating zones of environmental similarity (Erickson *et al.*, 2004; Johnson *et al.*, 2010; Miller *et al.*, 2011). However, the environmental categorization offered by an ecoregion may be too broad to detect certain limiting environmental factors and adaptive patterns (Parra-Quijano *et al.*, 2012a). On the other hand, the sensitivity to limiting environmental factors is different for each species. Therefore, a species-specific approach is necessary, such as that used in the study by Marinoni *et al.* (2021) for the conservation of plant genetic resources of species based on Ecogeographic Land Characterization (ELC) maps. ELC maps can be created for a species or a group of phylogenetically related species, using ecogeographic variables closely related to the distribution of the target species and delineating different adaptive scenarios within a given territory (García *et al.*, 2017; Parra Quijano *et al.*, 2012a). These characteristics make ELC maps a very appropriate provisional STZ system as indicated by Thomas *et al.* (2018).

In any case, a STZ of any type must be positively validated as a method to ensure the local adaptation of the transferred germplasm, through techniques such as common gardens (McKay *et al.*, 2005; Miller *et al.*, 2011).

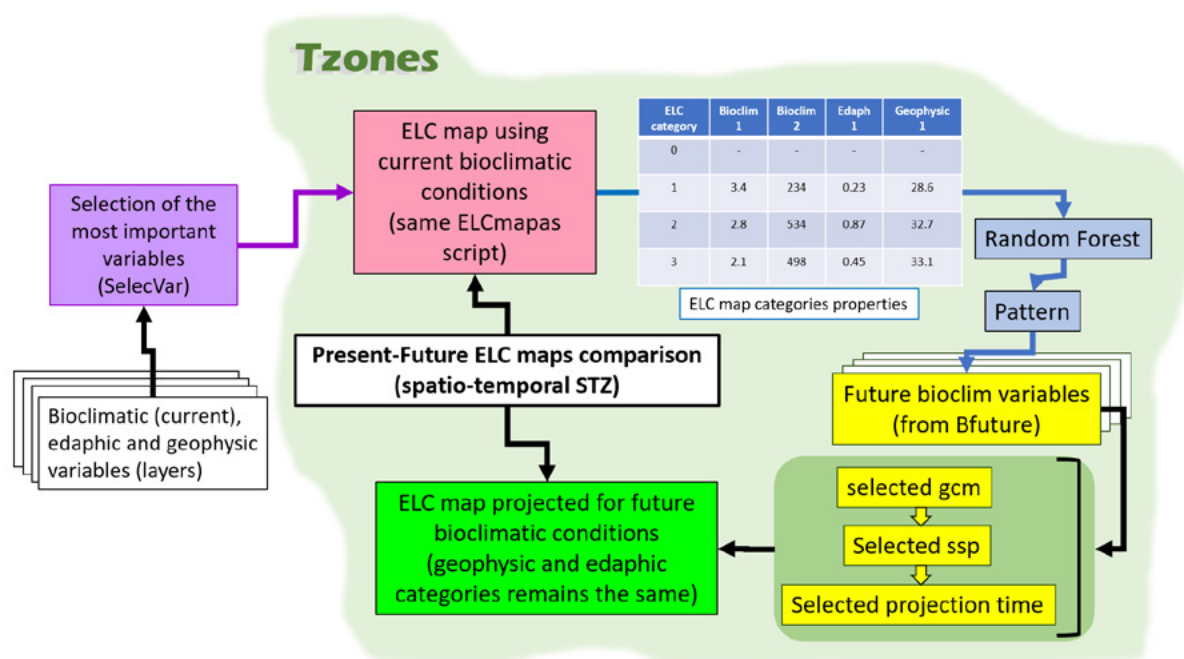
### 17.3. Spatio-temporal approximation in STZ

The first studies based on STZs were limited to present-time situations, considering only transfer in terms of two geographic locations (Withrow-Robinson & Johnson 2006; Bower *et al.*, 2014). In recent years, the provisional STZ approach has acquired a temporal dimension in addition to the previously mentioned spatial dimension. This new dimension could be key regarding conditions of rapid climate change, the need for material with appropriate adaptation, and ecological restorations in the future. The spatio-temporal approach consists of producing provisional STZs under present conditions and then projecting them into the future. This approach considers the environmental similarity not only between two sites in the present but also between one site in the present and another in the future. Therefore, spatial limits for seed transfer are established and defined for different periods (Potter & Hargrove 2012; Havens *et al.*, 2015; Richardson & Chaney 2018; Shryock *et al.*, 2018).

This additional dimension was introduced by Potter and Hargrove (2012), establishing the correspondence between present and future conditions under climate change conditions for 30,000 global ecoregions. However, in that study the authors carried out a generalist approach, assuming that a single global ecoregion map can be valid for any plant species. They considered that a single map of ecoregions is capable of correctly delineating adaptive scenarios for a wide range of species. However, local adaptation can vary significantly from species to species since the processes of natural selection and gene flow do not operate uniformly. This implies that provisional STZs must be developed at the species level, considering only the environmental variables of greatest importance regarding the adaptation of said species in the delineation of each zone (Johnson *et al.*, 2004; Kramer *et al.*, 2015).

## 17.4. How does Tzones work?

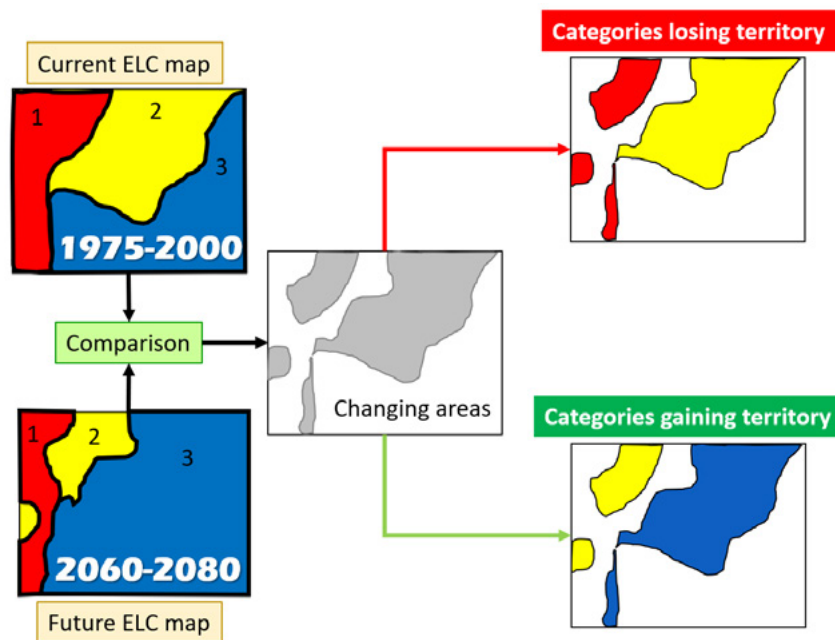
Tzones tool from CAPFITOGEN3 facilitates the identification of provisional and spatio-temporal STZs based on ELC maps created with current bioclimatic variables. The maps are projected into the future using the Random Forest algorithm and bioclimatic variables projected into potential future climate change scenarios. These potential scenarios are defined by global circulation models (GCM), shared socioeconomic pathways (SSP), and periods (years).



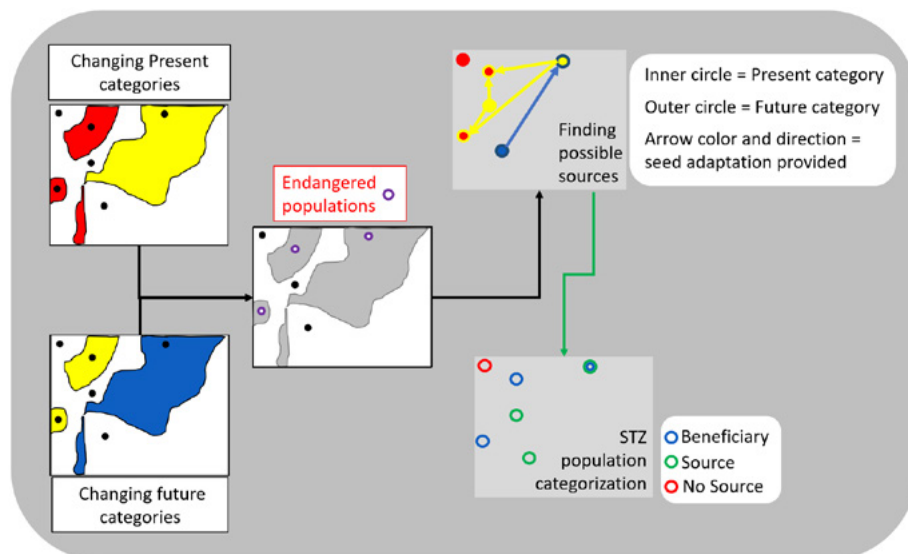
**Figure 58.** Procedure used by Tzones to create ELC maps for current conditions and then project them into the future.

In addition to the ELC map projection, Tzones determines the populations that could be affected by climate change in the future. This can be done since the ELC map category where populations currently occur is undergoing enough bioclimatic changes to modify the projected ELC map category. Finally, Tzones identifies the populations that, in the present, could serve as a source of germplasm for restorations of the affected populations in the future. The tool

determines these populations based on the abiotic environmental similarity that is assumed between two sites that belong to the same category of an ELC map (Figs. 59 and 60).



**Figure 59.** Example of the projection of a future ELC map (period 2070) and its comparison with the current map. This allows the identification of changing areas, where some categories can gain territory and others lose it.



**Figure 60.** Results of Tzones tool. Based on population occurrence in changing areas in the comparison between present-future ELC maps, the following populations are identified: a) future endangered populations that would potentially be beneficiaries of germplasm, b) populations that are sources of germplasm in the present to benefit those that will be endangered in the future, and c) populations that would not be affected and would not be sources of germplasm.

## 17.5. Using Tzones tool

Once CAPFITOGEN3 *local mode* tools have been installed or CAPFITOGEN3 *on server mode* has been accessed and Tzones tool has been selected, a series of parameters must be specified by the user.

### 17.5.1 Initial parameters defined by the user

#### 17.5.1.1 Parameter: ruta (only for local mode)

Explanation: Path where CAPFITOGEN tools have been copied or are to be found. Note: use / (slash) instead of \ (backslash) when indicating the path of the folder. For example, F:/CAPFITOGEN3, C:/CAPFITOGEN3, D:/MisHeramientas/CAPFITOGEN3, etc.

#### 17.5.1.2 Parameter: elcready

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if Tzones tool has been previously run for the same geographic space, occurrence/collection points and ecogeographic variables. Previous use of the tool generates a file called Productos.RData in the 'resultados' folder. If you want to use the same ELC map that was previously generated, but projecting it into other future scenarios, it is not necessary to generate the current map again. Simply indicate this option as TRUE or ✓. If the tool has not been used previously the parameter 'elcready' must be set as FALSE or □ whenever Tzones tool is going to be run for the first time for an area, a set of variables or occurrences/collection sites.

#### 17.5.1.3 Parameter: pais

Explanation: Select the country for which you wish to build the ELC map. In the drop-down list, there are more than 160 countries and some sub-continental, global, and custom coverages (from rLayer tool).

#### 17.5.1.4 Parameter: resol1

Explanation: Select the degree of resolution you wish to use to generate the map. Note that 1x1 km offers greater resolution but requires greater computing capacity and takes far longer than 5x5 km, particularly in countries with a large land mass. High-resolution maps (i.e., 1x1 km) composed of more than 100,000 cells (countries with more than 100,000 km<sup>2</sup>) can generate processing problems for some methods of clustering and determination of the optimal number of groups such as 'cluster'.

#### 17.5.1.5 Parameter: bioclimv

Explanation: List (*local mode*) or select (*on server mode*) the bioclimatic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter

script, in a line blocked with # (line announced like this: '#Complete list of bioclimatic variables'); copy the names and paste them separated by a semicolon (;). You will also find a blocked line with the 19 bioclim variables ready to use; simply remove the initial # symbol. In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'bioclimv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 17.5.1.6 Parameter: edaphv

Explanation: List (*local mode*) or select (*on server mode*) the edaphic variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of edaphic variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'edaphv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 17.5.1.7 Parameter: geophysv

Explanation: List (*local mode*) or select (*on server mode*) the geophysical variables you want to analyse. To select multiple variables in *local mode*, look for the names of the variables in the complete list that appears in the parameter script, in a line blocked with # (line announced like this: '#Complete list of geophysical variables'); copy the names and paste them separated by a semicolon (;). In *on server mode*, select each variable of interest by clicking on its name. Once selected, the name of the variable will appear in the box in front of parameter 'geophysv'. To know the codes, names, and brief descriptions of the variables, check the 'Variables names - Nombres de variables.xlsx' file downloadable from <http://www.capfitogen.net/es/Variables-names-Nombres-de-variables.xlsx> and also available in the installation of CAPFITOGEN3 *local mode*.

#### 17.5.1.8 Parameter: latitud

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) to include the latitude (Y) as a geophysical variable to be analysed.

#### 17.5.1.9 Parameter: longitud

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) to include the longitude (X) as a geophysical variable to be analysed.

#### 17.5.1.10 Parameter: *ssp* (previously *rcp*)

Explanation: Select the Representative Concentration Pathway (RCP) for the projected bioclimatic data you have available for the country/region ('pais' parameter) in the 'rdatamapsf2' folder. RCPs represent the greenhouse gas concentration scenario for your 'predicted bioclimatic data'. There are four greenhouse gas concentration trajectories adopted by the IPCC: RCP2.6 (here appears as 26), RCP4.5 (45), RCP6 (60), and RCP8.5 (85). These options are related to radiative values in the year 2100 compared to preindustrial values (+2.6, +4.5, +6.0, and +8.5 W/m<sup>2</sup>, respectively). There are four available SSP: 126 (ssp1, 2.6), 245 (ssp2, 4.5), 370 (ssp3, 7.0) and 585 (ssp5, 8.5).

#### 17.5.1.11 Parameter: *gcm*

Explanation: Applies only if 'temp' = 'future'. Select the Global Climate Model (GCM, also known as General Circulation Model) for the projected bioclimatic data you have available for the country/region (parameter 'pais') in the 'rdatamapsf' folder. These models simulate future weather scenarios according to the assumed atmospheric concentration of greenhouse gasses (rcp/ssp). If you want to know more about GCM, please visit <http://goo.gl/4XhU6g> or <http://goo.gl/2VUuRP>.

#### 17.5.1.12 Parameter: *proy*

Explanation: Select the period for which you want to project current ELC maps. Option 50 represents the value/year 2050 (average for 2041-2060) and option 70 represents the year 2070 (average for 2061-2080). Make sure you have the bioclimatic information projected in the future for parameters *ssp*, *gcm*, and *proy*. The information, which should be according to the work frame (parameter 'pais'), must be in the *rdatamapsf* folder (in *local mode* version). Consider reviewing the function and process associated with *Bfuture* to appropriately set up these three parameters (*ssp*, *gcm*, and *proy*).

#### 17.5.1.13 Parameter: *maxg*

Explanation: Indicate the maximum number of clusters per component (bioclimatic, geophysical, and edaphic) that you wish to allow (the larger the number, the more categories on the map). We recommend values lower than or equal to five, otherwise, ELC maps of more than 125 categories can be generated.

#### 17.5.1.14 Parameter: *metodo*

Explanation: Select one of the methods offered to generate the clusters with an objective determination of the optimal number of clusters. The six methods described in this chapter are available by the terms: 'kmeansbic', 'medoides', 'elbow', 'calinski', 'ssi' and 'bic'.



#### 17.5.1.15 Parameter: *iterat*

Explanation: Applies only when you have selected 'calinski' or 'ssi' as methods in parameter 'metodo'. Indicate in this field the number of iterations used to generate K-means clusters to calculate 'calinski' or 'ssi' criteria.

#### 17.5.1.16 Parameter: *iteratf*

Explanation: This parameter indicates the number of iterations or permutations to be used by the Random Forest algorithm with which the present ELC map is projected with future bioclimatic conditions/variables under an SSP, GCM, and future period (years).

#### 17.5.1.17 Parameter: *transfer*

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if you want to perform a germplasm transfer zone analysis. This procedure assumes that populations occurring in changing categories between current and future ELC maps would be at risk of poor adaptation. Thus, they may need to be strengthened through the introduction of germplasm from populations that currently occur in future ELC categories of affected populations. A table with occurrence data (whose name must be indicated in parameter 'pasaporte') is required. This table includes populations that could be at risk of disappearing due to climate change and populations that could serve as a germplasm source in the present to prevent their disappearance.

#### 17.5.1.18 Parameter: *pasaporte*

Explanation: For *local mode*, enter the name of the file containing the passport table in text format, remembering to add the file extension (.txt). For example, if the file is named 'table', you should enter 'table.txt'. Please remember to save this file first in the 'Pasaporte' folder, which is part of the set of folders that make up the CAPFITOGEN directory. Data in the 'Pasaporte' folder may be in subfolders that must be reflected in this field. For example, if your table is called 'table.txt' and is in a folder called 'Stuberosum' that is inside 'Pasaporte', then Stuberosum/table.txt should appear in this field, always using / (slash) instead of \ (backslash) in the path description. For *on server mode*, you only have to upload the .txt file or select it from those that have been previously uploaded and that are in the User's Files and Results area.

#### 17.5.1.19 Parameter: *tecogaps*

Explanation: Select this option (TRUE in *local mode* or ✓ in *on server mode*) if your passport table includes column 'GAPTYPE' with valid values (as it has been analysed by the Representa tool) to filter those occurrences corresponding to priority ecogeographic gaps. Using the gaps detected in the previous analysis by Representa, Tzones identifies as priorities for collection the populations that are both sources and ecogeographic/ spatial gaps.

### 17.5.1.20 Parameter: *ttresh*

Explanation: Applies only if ‘*tecogaps*’ has been set as TRUE in *local mode* or ✓ in *on server mode*. Enter a value for parameter ‘*ttresh*’ that will be used as a threshold value for GAPTYPE (from the scale of 1 to 15 generated by the Representa tool) to determine which occurrences are considered priority ecogeographic gaps. This value is usually set to 4 (see chapter for Representa tool).

### 17.5.1.21 Parameter: *resultados* (only for *local mode*)

Explanation: Enter the path of the folder where you wish the results of the analysis to be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados, D:/MisHerramientas/Resultados, etc. Current ELC maps will be saved in this folder.

### 17.5.1.22 Parameter: *resultados2* (only for *local mode*)

Explanation: Enter the path of the folder where you wish the results of the future analysis to be saved. Note: use / (slash) instead of \ (backslash). For example, C:/Resultados2, D:/MisHerramientas/Resultados2, etc. The path here must be different from that of parameter ‘*resultados*’.

## 17.6. Results of Tzones

After defining all the parameters and paths (for *local mode*) that Tzones requires, the analysis process of the tool will start by clicking on the ‘Run’ button (*local mode* in RStudio) or ‘Start’ (*on server mode*).

After some time, which may vary due to the type of analysis requested, Tzones will save the results in the paths and folders specified in ‘*resultados*’ and ‘*resultados2*’ (in *local mode*), or in the User’s Files and Results area in *on server mode*.

### 17.6.1 Folder indicated in parameter ‘*resultados*’

This folder contains the same products of the execution of the ELCmaps tool, with the addition of the *Producto.RData* file. In this file, the necessary elements of this ELC map (created with current bioclimatic variables) are saved for future utilization in its future projection.

### 17.6.2 Folder indicated in parameter ‘*resultados2*’

All the results regarding the projection of the ELC map under certain future and climate change scenarios will be saved in this folder. Also, the analysis of the changes between the current and future ELC map, the changes of catego-

ry for the occurrence/collection sites and, therefore, the system of seed source-recipient or beneficiary populations (spatio-temporal seed transfer zones) will also be saved in this folder.

**17.6.2.1 'TransferZoneAnalysis.txt/xls':** This table verifies for each presence/collection if the ELC category has changed or remains the same. It also identifies the populations that are endangered by said change and the possible sources of germplasm for future restorations. Column 'PopEndangered' identifies the populations for which the ELC category will change (indicating the number of the category where this population currently grows), and column 'Sources' indicates (with a value of 1) those populations that could be collected in the present to restore endangered populations in the future.

**17.6.2.2 'stats\_change\_table.txt/xls':** This table shows the percentage of change (expressed as those cells that change against the total cells of each category) for each category of the ELC map compared to the ELC map projected in the future (column 'Freq').

**17.6.2.3 'stats\_change\_table.txt/xls':** This table shows the characteristics of the ELC map projected into the future in terms of each ecogeographic variable used to generate the current ELC map.

**17.6.2.4 'mapa\_sspXXX\_elcF\_pais.tif/grd':** Corresponds to the future ELC map (projected).

**17.6.2.5 'mapa\_sspXXX\_elcF\_DIVA\_pais.grd':** This map is a copy of the map mentioned above. DIVA-GIS changes some characteristics of the map when opening it, so this file is generated in case the user needs an original copy.

**17.6.2.6 'mapa\_sspXXX\_bioclimaticoF\_pais.tif/grd':** This map corresponds to the projection of the map of current bioclimatic data. The projection is combined with the geophysical and edaphic maps to generate the ELC map.

**17.6.2.7 'map\_change\_pres\_pais.tif/grd' and 'map\_change\_fut\_pais.tif/grd':** These maps exclusively show the changing categories and the result of the comparison between present-future ELC maps. Both maps contain the same changing areas, only with the difference that the map 'map\_change\_pres\_pais' shows the areas with their current values and 'map\_change\_fut\_pais' shows the areas with the values they have in the future map.

## 17.7. References

Bates, C. G. 1930. The frost hardiness of geographic strains of Norway pine. *Journal of Forestry* 28(3): 327-333.

Bower, A. D., Clair, J. B. S., Erickson, V. 2014. Generalized provisional seed zones for native plants. *Ecological Applications* 24(5): 913-919.

Cevallos, D., Bede-Fazekas, Á., Tanács, E., Szitár, K., Halassy, M., Kövendi-Jakó, A., Török, K. 2020. Seed transfer zones based on environmental variables better reflect variability in vegetation than administrative units: evidence from Hungary. *Restoration Ecology* 28(4): 911-918.

Chuine, I. 2010. Why does phenology drive species distribution? *Philosophical Transactions of the Royal Society B* 365: 3149-3160.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J. 2007. Random forests for classification in ecology. *Ecology* 88(11): 2783-2792.

Dinerstein, E. D., Olson, A., Joshi, C., Vynne, N., Burgess, E., Wikramanayake, N., Hahn, S., Palminteri, P., Hedao, R., Noss, M., Hansen, H., Locke, E., Ellis, B., Jones, C., Barber, V., Hayes, R., Kormos, C., Martin, V., Crist, E., Sechrest, W., Price, L., Baille, J., Weeden, D., Suckling, K., Davis, C., Sizer, N., Moore, R., Thau, D., Birch, T., Potapov, P., Turubanova, S., Tyukavina, A., Souza, N., Pintea, L., Brito, J., Llewellyn, O., Miller, A., Patzelt, A., Ghazanfar, S., Timberlake, J., Klozer, H., Shenan-Farpon, Y., Kindt, R., Barnekow, J., van Breugel, P., Graudal, L., Vogé, M., Al-Shammari, K., Saleem, M. 2017. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience* 67(6): 534-545.

Doherty, K. D., Butterfield, B. J., Wood, T. E. 2017. Matching seed to site by climate similarity: techniques to prioritize plant materials development and use in restoration. *Ecological Applications* 27(3): 1010-1023.

Durka, W., Michalski, S. G., Berendzen, K. W., Bossdorf, O., Bucharova, A., Hermann, J. M., Holzel, N., Kollmann, J. 2017. Genetic differentiation within multiple common grassland plants supports seed transfer zones for ecological restoration. *Journal of Applied Ecology* 54(1): 116-126.

Erickson, V. J., Mandel, N. L., Sorensen, F. C. 2004. Landscape patterns of phenotypic variation and population structuring in a selfing grass, *Elymus glaucus* (blue wildrye). *Canadian Journal of Botany* 82(12): 1776-1789.

Feeley, K. J., Silman, M. R. 2009. Extinction risks of Amazonian plant species. *Proceedings of the National Academy of Sciences* 106(30): 12382-12387.

Fowells, H. A. 1949. Cork oak planting tests in California. *Journal of Forestry* 47(5): 357-365.

García, R. M., Parra-Quijano, M., Iriondo, J. M. 2017. A multispecies collecting strategy for crop wild relatives based on complementary areas with a high density of ecogeographical gaps. *Crop Science* 57(3): 1059-1069.

Germiño, M. J., Moser, A. M., Sands, A. R. 2019. Adaptive variation, including local adaptation, requires decades to become evident in common gardens. *Ecological Applications* 29(2): e01842.

Gibson, A., Nelson, C. R. 2017. Comparing provisional seed transfer zone strategies for a commonly seeded grass, *Pseudoroegneria spicata*. *Natural Areas Journal* 37(2): 188-199.

Hamann, A., Gylander, T., Chen, P. Y. 2011. Developing seed zones and transfer guidelines with multivariate regression trees. *Tree Genetics & Genomes* 7(2): 399-408.

- Hanson, J. O., Rhodes, J. R., Riginos, C., Fuller, R. A. 2017. Environmental and geographic variables are effective surrogates for genetic variation in conservation planning. *Proceedings of the National Academy of Sciences* 114(48): 12755-12760.
- Harris, J. A., Hobbs, R. J., Higgs, E., Aronson, J. 2006. Ecological restoration and global climate change. *Restoration Ecology* 14(2): 170-176.
- Havens, K., Vitt, P., Still, S., Kramer, A. T., Fant, J. B., Schatz, K. 2015. Seed sourcing for restoration in an era of climate change. *Natural Areas Journal* 35(1): 122-133.
- Higgs, E. S. 1997. What is good ecological restoration? *Conservation Biology* 11(2): 338-348.
- IPCC. 2013. IPCC Fifth Assessment Report (AR5). IPCC s. 10-12.
- Johnson, G. R., Sorensen, F. C., St Clair, J. B., Cronn, R. C. 2004. Pacific northwest forest tree seed zones a template for native plants? *Native Plants Journal* 5(2): 131-140.
- Johnson, R. C., Erickson, V. J., Mandel, N. L., St Clair, J. B., Vance-Borland, K. W. 2010. Mapping genetic variation and seed zones for *Bromus carinatus* in the Blue Mountains of eastern Oregon, USA. *Botany* 88(8): 725-736.
- Jombart, T., Collins, C. 2015. A tutorial for discriminant analysis of principal components (DAPC) using adegenet 2.0.0. London: Imperial College London, MRC Centre for Outbreak Analysis and Modelling.
- Kramer, A. T., Havens, K. 2009. Plant conservation genetics in a changing world. *Trends in Plant Science* 14(11): 599-607.
- Kramer, A. T., Larkin, D. J., Fant, J. B. 2015. Assessing potential seed transfer zones for five forb species from the Great Basin Floristic Region, USA. *Natural Areas Journal* 35(1): 174-188.
- McKay, J. K., Christian, C. E., Harrison, S., Rice, K. J. 2005. "How local is local?"—a review of practical and conceptual issues in the genetics of restoration. *Restoration Ecology* 13(3): 432-440.
- Marinoni, L., Parra Quijano, M., Zabala, J.M., Pensiero, J.F., Iriondo, J.M. 2021. Spatio-temporal seed transfer zones as an efficient restoration strategy in response to climate change. *Ecosphere*, in press.
- Miller, S. A., Bartow, A., Gisler, M., Ward, K., Young, A. S., Kaye, T. N. 2011. Can an ecoregion serve as a seed transfer zone? Evidence from a common garden study with five native species. *Restoration Ecology* 19(201): 268-276.
- Omernik, J. M., Griffith, G. E. 2014. Ecoregions of the conterminous United States: evolution of a hierarchical spatial framework. *Environmental management* 54(6): 1249-1266.

Parra Quijano, M., Iriondo, J. M., Torres, E. 2012a. Ecogeographical land characterization maps as a tool for assessing plant adaptation and their implications in agrobiodiversity studies. *Genetic Resources Crop Evolution* 59: 205-217.

Parra Quijano, M., Iriondo, J. M., Torres, E. 2012b. Applications of ecogeography and geographic information systems in conservation and utilization of plant genetic resources. *Spanish Journal of Agricultural Research* 10(2): 419-429.

Parra Quijano, M., Iriondo, J.M., Frese, L., Torres, E. 2012c. Spatial and ecogeographic approaches for selecting genetic reserves in Europe. In: Maxted, N., Dulloo, M.E., Ford-Lloyd, B.V., Frese, L., Iriondo J., and Pinheiro de Carvalho M.A.A. (eds.) *Agrobiodiversity conservation: securing the diversity of crop wild relatives and landraces*. CABI, Wallingford, UK.

Parra Quijano, M. 2016. Tools CAPFITOGEN Program to Strengthen Capabilities in National Plant Genetic Resources Programs in Latin America.

Peeters, J. P., Wilkes, H. G., Galwey, N. W. 1990. The use of ecogeographical data in the exploitation of variation from gene banks. *Theoretical and Applied Genetics* 80(1): 110-112.

Potter, K. M., Hargrove, W. W. 2012. Determining suitable locations for seed transfer under climate change: a global quantitative method. *New Forests* 43: 581-599.

Richardson, B. A., Chaney, L. 2018. Climate-based seed transfer of a widespread shrub: population shifts, restoration strategies, and the trailing edge. *Ecological Applications* 28(8): 2165-2174.

Shryock, D., Defalco, L. A., Esque, T. C. 2018. Spatial decision-support tools to guide restoration and seed-sourcing in the Desert Southwest. *Ecosphere* 9(10): 1-19.

Thomas, E., Alcazar, C., Moscoso L. G., Vásquez A., Osorio L. F., Salgado-Negrete, B., Gonzalez, M., Parra-Quijano, M., Bozzano, M., Loo, J., Jalonen, R., Ramírez, W. 2017. The importance of species selection and seed sourcing in forest restoration for enhancing adaptive capacity to climate change: Colombian tropical dry forest as a model. *The Lima declaration on biodiversity and climate change: contributions from science to policy for sustainable development* (89): 122-132.

Williams, M. I., Dumroese, R. K. 2013. Preparing for climate change: Forestry and assisted migration. *Journal of Forestry* 111(4): 287-297.

Withrow-Robinson, B. A., Johnson, R. 2006. Selecting native plant materials for restoration projects: ensuring local adaptation and maintaining genetic diversity Oregon State University. URL: <https://ir.library.oregonstate.edu/downloads/g732d9349>









# 18 | Frequent Errors

The following list shows many of the error messages that you can find in *local mode* or in *on server mode*. For *local mode*, errors are shown in red as they would appear in the RStudio window (RStudio window shown in Fig. 2-C). For *on server mode*, they are shown a text file in the folder where results are stored in the User's Files and Results area.

- **Error message:** An error occurred: Error in readChar(con, 5L, useBytes = TRUE) : conexiónCalls cannot be opened: source -> withVisible -> eval -> eval -> load -> readChar

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** Any tool

**Solution(s):** This error, the most common of all, usually corresponds to incorrect parameter entries or indications. Sometimes some part of the script changes (according to the tool or the part of the script the change occurs), but in both cases, it always displays the following: **connectionCalls cannot be opened**. For example, in ELCmaps, this error may appear if a cell resolution of 10x10 km is indicated for a country like Cuba in this tool (the tool will look for this resolution in CAPFITOGEN3/rdatamaps/cuba/ and it will not find it, because it does not exist). This can also happen when a wrong path is indicated to locate tools or passport tables, etc. To avoid this problem, check each parameter confirming that the values are correct.

- **Error message:** An error occurred: Error in library(package name) : there is no package called 'package name'Calls: source -> withVisible -> eval -> eval -> library

**Mode in which the error occurs:** Local

**Tool in which the error occurs:** Any tool, as this is an installation issue.

**Solution(s):** The error indicates that one of the R packages the tool requires was not properly installed, which is why R cannot find it. Make sure that the structure of folders and files for the tools is not located in the root directory (for example in K:/). If this is the case, create a folder (usually named CAPFITOGEN3) in the root directory, then cut and paste the entire folder set into the new folder created. Then reinstall the tools. If this option does not work, try installing the package manually. To do this:

- The error code gives the name of the uninstalled package, exactly where it says 'package name' in the example. Use the package name to open the 'packages' folder in the set of CAPFITOGEN3 folders and files. Here you will find a series of '.zip' files with different names. One of these files corresponds to the package name and is accompanied by numbers that refer to the version. Copy the file name completely and include the .zip extension.
- Open RStudio and type the following command in the 'R console': `Install.packages('X:/CAPFITOGEN3/packages/nombreadivopaque.zip')` where X refers to the drive where CAPFITOGEN tools are located (change this letter accordingly). Paste the file name copied in step 1 where it says 'nombreadivopaque.zip'. Then press 'enter'.
- The program will install the package and when it has finished, the following notification will appear. package 'sp' successfully unpacked and MD5 sums checked.
- Make sure the package has been successfully installed by typing: `library('package name')` where 'package name' is the name of the package as it appears in the error notice (with neither the version code nor the .zip extension). Then press 'enter'. The following notification will appear: Lost warning notices package 'cluster' was built under R version 2.15.3.
- Try the tool again. The same error may appear again but for a different package. If so, repeat the operation

until the error notices cease to appear. Such errors tend to be unusual since the installation system was improved, but they do occur occasionally, particularly with Windows 8.

- Error message:** An error occurred: An error occurred: Error: unable to locate a vector of X.X Gb  
**Mode in which the error occurs:** Local and on server  
**Tool in which the error occurs:** Various  
**Solution(s):** This problem is related to the size of the matrices managed by R. It can usually be solved by reducing the resolution of the maps. If the error occurs in ELCmapas, change the method to determine the optimal number of groups or increase the cell size in the resol1 parameter. This error may also appear in GEOQUAL due to an error in the contents of the passport tables, specifically when duplicates occur in the field ACCENUMB. This field unequivocally identifies each accession and, thus, a single duplicate can generate an error message. The solution is to check that there are no duplicates in the table. If there are any, assign each duplicate accession a unique number or code.
- Error message:** An error occurred: Error in sample.int(m, k) : first argument InvalidoCalls: source... withVisible -> eval -> eval -> kmeans -> sample.int  
**Mode in which the error occurs:** Local and on server  
**Tool in which the error occurs:** ELCmapas  
**Solution(s):** This means that a variable is constant for that region or country and that when it is standardized, it produces a table of 0 rows which generates an error in Kmeans (elbow method). It can be solved by deselecting the variable causing the problem. Please note that this variable usually corresponds to soil variables, especially in small countries. For example, the variable 'depth' often creates this problem. Using minimum rainfall variables in dry countries also tends to produce this problem.
- Error message:** An error occurred: Error in clara(sdata, k,...) : x is not a numeric dataframe or matrix.Calls: source -> withVisible -> eval -> eval -> pamk -> clara  
**Mode in which the error occurs:** Local and on server  
**Tool in which the error occurs:** ELCmapas  
**Solution(s):** This means that a variable is constant for that zone and that when it is standardized, it produces a table of 0 rows which generates an error in medoides. The solution is the same as for No. 2.
- Error message:** An error occurred: Error in kmeans(edaph[,-1], centers = i) : more cluster centers than distinct data points.Calls: source -> withVisible -> eval -> eval -> kmeans  
**Mode in which the error occurs:** Local and on server  
**Tool in which the error occurs:** ELCmapas  
**Solution(s):** This means that the maximum number of groups entered is lower than the optimum target number determined by the elbow method. Repeat the operation with a lower number of groups.
- Error message:** An error occurred: Error: 'ecogeot' object not found  
**Mode in which the error occurs:** Local and on server  
**Tool in which the error occurs:** ECOGEO

**Solution(s):** Select the geophysv option if selecting geophysical variables.

- Error message:** An error occurred: Error in validObject (.Object) : invalid class "SpatialPoints" object: bbox should never contain infinite valuesCalls: source ... SpatialPoints -> new -> initialize -> initialize -> validObject

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** Representa

**Solution(s):** Review the text file called 'process\_info.txt' in the 'Error' folder in the CAPFITOGEN tools set of folders and files. The bottom line of the text file may read, 'WARNING! failed to delete all FE records as data from other banks considered not missing'. This indicates that the tool has run out of data from external sources because all the contributions have been sourced from 'germplasm banks'. When instructed to assume that these are not missing, an error occurs as there is no data left to analyze. Remove the option for external sources or allow Representa to use data from other banks as missing.
- Error message:** An error occurred: Error in dist(x[ss[[i]],], method = metric, ...) : longitude vectors not allowed negativaCalls: source ... withVisible -> eval -> eval -> pamk -> distcritmulti -> dist

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** ELCmapas

**Solution(s):** This error appears when the country or region is very large, the resolution is high (a smaller cell size), and the tool is asked to determine the optimum number of 'medoides' clusters. The first solution is to rerun the analysis using the elbow method. If, regardless of this, another error is generated, use a lower resolution (larger cell size).
- Error message:** An error occurred: Error in merge.data.frame(as.data.frame(x), as.data.frame(y),...) : longitude vectors are not allowed negativaCalls: source ... merge -> merge.default -> merge -> merge.data.frame

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** ELCmapas

**Solution(s):** The error persists because the matrices generated are so large that the elbow method to determine the optimum number of clusters cannot manage them. The solution is to use a lower resolution (greater cell size).
- Error message:** An error occurred: Error in .checkNumericCoerce2double(obj) : cannot retrieve coordinates from non-numeric elementsCalls: source ... coordinates -> .local -> do.call -> .checkNumericCoerce2double

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** GEOQUAL

**Solution(s):** Error in coding the coordinates or preparing the passport table. In the first case, correct the coordinates manually in Excel and save the file in tab-delimited text format. In the second case, the order of the variables is wrong, which is why the columns corresponding to the coordinates are misplaced. Follow the order of the variables exactly according to the format specified and do not add columns or change their order.

- **Error message:** An error occurred: Error in apply(x, 2, fun2) : dim (X) must have a positive lengthCalls: source ... extract -> .xyValues -> .xyvBuf -> lapply -> FUN -> apply

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** This may occur when using radial extraction tools.

**Solution(s):** This error may occur when the user requests a radial extraction using a radius that is too small (parameter `tamp`) for the cell size or ecogeographic variable resolution (parameter `'resol1'`). For example, if you request a radial extraction of 1 km using cell resolutions of 10x10 km approx. (5 arc-min). This will produce extraction values of zero and generate an error. Try using larger radii, ensuring they are greater than the size of the side of each cell, and/or use a higher resolution. For example, if working with a radial extraction of 1 km, change `'Celdas 5x5 km approx. (2.5 arc-min)'` to `'Celdas 1x1 km approx. (30 arc-sec)'` to solve the problem. If this does not work, try using specific extractions.

- **Error message:** An error occurred: Error in 'colnames<-'('tmp\*', value = "ACCENUMB") : the 'names' [1] attribute must have the same length as the vector [0]Calls: source -> withVisible -> eval -> eval -> colnames <-

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** This may occur with tools where the user needs to enter passport data.

**Solution(s):** The error message may occur when, in parameter `'pasaporte'`, the user indicates a passport table with the wrong number of columns. This may be due to the accidental deletion of a column, or because the tool expects additional columns that are not included. This can occur with `ColNucleo`, which expects the additional `'AVAILAB'` column. It can also occur if, under parameter `'geoqual'`, the user indicates that the table has four extra columns containing the results of the `GEOQUAL` analysis when in fact it does not. Check the contents of the passport table you are entering and use the parameter `'geoqual'` accordingly.

- **Error message:** An error occurred: Error in if (any(puntosorig\$DECLATITUDE >= 90 puntosorig\$DECLATITUDE <= : value absent where TRUE/FALSE is necesarioCalls: source -> withVisible -> eval -> eval or An error occurred: Error in if (any(puntosorig\$DECLONGITUTE >= 180 | puntosorig\$DECLONGITUDE <= : value absent where TRUE/FALSE is necesarioCalls: source-> withVisible-> eval-> eval

**Mode in which the error occurs:** Local and on server

**Tool in which the error occurs:** This may occur with tools where the user needs to enter passport details.

**Solution(s):** There is an error in at least one of the accession's coordinates, which may be due to mistakes in coding the coordinates or because the coordinate field is empty or NA. To solve the problem in the first case, check the full six-figure code of the coordinates and ensure these correspond to the FAO/Bioversity 2012 format. For decimal values, these are between -90 and 90 for `DECLATITUDE` and between -180 and 180 for `DECLONGITUDE`. In the second case (empty or NA fields), this may be due to the emergence of 'ghost' accessions, which are formed when the passport table is created in Excel. This table has extra rows that unfortunately cannot be easily identified as they are blank and only appear when you export the table in text format. The system interprets them as accessions because they occupy a row, but as they have neither data nor coordinates, this generates an error.

- **Error message:** An error occurred: Error in CRS(as.character(projection(crs))) : projection not namedCalls: source ... .rasterFromRasterFile -> raster -> raster -> .local -> CRS

**Mode in which the error occurs:** Local and on server

**Mode in which the error occurs:** This may occur in any of the tools that use ELC maps such as Representa, ColNucleo, or FIGS\_R.

**Solution(s):** The ELC map produced by the ELCmapas tool that was copied and pasted in the path CAPFITOGEN2/ ELCmapas was edited and the coordinate system was altered during this process. Thus, the R Raster package cannot read it. Please use an unedited ELC map, copy it and paste it into the path CAPFITOGEN2/ ELCmaps, and with that the problem should disappear. For this purpose, ELCmapas tool produces two copies of the map, one to be edited in DIVA-GIS and the other must be preserved without editing to be used later.









# 19 | Annexes

## 19.1. Names of variables available in CAPFITOGEN3, as they are used in *local mode* and in *on server mode*

ID	CODE	VARIABLE ( <i>Local mode</i> )	VARIABLE ( <i>On server mode</i> )	COMPONENT
1	alt	elevación	Elevation	Geophysical
2	aspect	Orientacion	Aspect	Geophysical
3	bio_1	Temp prom anual	Annual mean temp	Bioclimatic
4	bio_10	Temp prom cuarto mas cálido	Mean temp warmest quarter	Bioclimatic
5	bio_11	Temp prom cuarto mas frio	Mean temp coldest quarter	Bioclimatic
6	bio_12	Prec anual	Annual prec	Bioclimatic
7	bio_13	Prec mes mas humedo	Prec wettest month	Bioclimatic
8	bio_14	Prec mes mas seco	Prec driest month	Bioclimatic
9	bio_15	Estacionalidad prec	Prec seasonality	Bioclimatic
10	bio_16	Prec cuarto mas humedo	Prec wettest quarter	Bioclimatic
11	bio_17	Prec cuarto mas seco	Prec driest quarter	Bioclimatic
12	bio_18	Prec cuarto mas calido	Prec warmest quarter	Bioclimatic
13	bio_19	Prec cuarto mas frio	Prec coldest quarter	Bioclimatic
14	bio_2	Rango prom temp diurnas	Mean temp diurnal range	Bioclimatic
15	bio_3	Isotermalidad	Isothermality	Bioclimatic
16	bio_4	Estacionalidad temp	Temp seasonality	Bioclimatic
17	bio_5	Max temp mes mas calido	Max temp warmest month	Bioclimatic
18	bio_6	Min temp mes mas frio	Min temp coldest month	Bioclimatic
19	bio_7	Rango temp anual	Temp annual range	Bioclimatic
20	bio_8	Temp prom cuarto humedo	Mean temp wettest quarter	Bioclimatic
21	bio_9	Temp prom cuarto seco	Mean temp driest quarter	Bioclimatic
22	eastness	Esticidad	Eastness	Geophysical
23	northness	Norticidad	Northness	Geophysical
24	prec_1	Prec prom 1	Prec mean 1	Bioclimatic
25	prec_10	Prec prom 10	Prec mean 10	Bioclimatic
26	prec_11	Prec prom 11	Prec mean 11	Bioclimatic
27	prec_12	Prec prom 12	Prec mean 12	Bioclimatic
28	prec_2	Prec prom 2	Prec mean 2	Bioclimatic
29	prec_3	Prec prom 3	Prec mean 3	Bioclimatic

## Continued

ID	CODE	VARIABLE ( <i>Local mode</i> )	VARIABLE ( <i>On server mode</i> )	COMPONENT
30	prec_4	Prec prom 4	Prec mean 4	Bioclimatic
31	prec_5	Prec prom 5	Prec mean 5	Bioclimatic
32	prec_6	Prec prom 6	Prec mean 6	Bioclimatic
33	prec_7	Prec prom 7	Prec mean 7	Bioclimatic
34	prec_8	Prec prom 8	Prec mean 8	Bioclimatic
35	prec_9	Prec prom 9	Prec mean 9	Bioclimatic
36	ref_depth	Profundidad	Depth	Edaphic
37	s_bs	Sat bases subsuelo	Subsoil base saturation	Edaphic
38	s_caco3	CaCO3 subsuelo	Subsoil CaCO3	Edaphic
39	s_caso4	Yesos subsuelo	Subsoil gypsum	Edaphic
40	s_cec_clay	CIC arcilla subsuelo	Subsoil CEC clay	Edaphic
41	s_cec_soil	CIC subsuelo gral	Subsoil CEC soil	Edaphic
42	s_clay	Arcilla en subsuelo	Subsoil clay fraction	Edaphic
43	s_ece	Salinidad subsuelo	Subsoil salinity	Edaphic
44	s_esp	Sodicidad subsuelo	Subsoil sodicity	Edaphic
45	s_gravel	Grava en subsuelo	Subsoil gravel	Edaphic
46	s_oc	Carbon org subsuelo	Subsoil org carbon	Edaphic
47	s_ph_h2o	pH subsuelo	Subsoil pH	Edaphic
48	s_ref_bulk	Densidad subsuelo	Subsoil bulk density	Edaphic
49	s_sand	Arena en subsuelo	Subsoil sand fraction	Edaphic
50	s_silt	Limo en subsuelo	Subsoil silt fraction	Edaphic
51	s_teb	Bases int subsuelo	Subsoil exchange bases	Edaphic
52	slope	Pendiente grados	Slope	Geophysical
53	t_bs	Sat bases suelo	Topsoil base saturation	Edaphic
54	t_caco3	CaCO3 suelo	Topsoil CaCO3	Edaphic
55	t_caso4	Yesos suelo	Topsoil gypsum	Edaphic
56	t_cec_clay	CIC arcilla suelo	Topsoil CEC clay	Edaphic
57	t_cec_soil	CIC suelo gral	Topsoil CEC soil	Edaphic
58	t_clay	Arcilla en suelo	Topsoil clay fraction	Edaphic
59	t_ece	Salinidad suelo	Topsoil salinity	Edaphic
60	t_esp	Sodicidad suelo	Topsoil sodicity	Edaphic

## Continued

ID	CODE	VARIABLE (Local mode)	VARIABLE (On server mode)	COMPONENT
61	t_gravel	Grava en suelo	Topsoil gravel	Edaphic
62	t_oc	Carbon org suelo	Topsoil org carbon	Edaphic
63	t_ph_h2o	pH suelo	Topsoil pH	Edaphic
64	t_ref_bulk	Densidad suelo	Topsoil bulk density	Edaphic
65	t_sand	Arena en suelo	Topsoil sand fraction	Edaphic
66	t_silt	Limo en suelo	Topsoil silt fraction	Edaphic
67	t_teb	Bases int suelo	Topsoil exchange bases	Edaphic
68	tmax_1	Temp max 1	Max temp 1	Bioclimatic
69	tmax_10	Temp max 10	Max temp 10	Bioclimatic
70	tmax_11	Temp max 11	Max temp 11	Bioclimatic
71	tmax_12	Temp max 12	Max temp 12	Bioclimatic
72	tmax_2	Temp max 2	Max temp 2	Bioclimatic
73	tmax_3	Temp max 3	Max temp 3	Bioclimatic
74	tmax_4	Temp max 4	Max temp 4	Bioclimatic
75	tmax_5	Temp max 5	Max temp 5	Bioclimatic
76	tmax_6	Temp max 6	Max temp 6	Bioclimatic
77	tmax_7	Temp max 7	Max temp 7	Bioclimatic
78	tmax_8	Temp max 8	Max temp 8	Bioclimatic
79	tmax_9	Temp max 9	Max temp 9	Bioclimatic
80	tmean_1	Temp prom 1	Mean temp 1	Bioclimatic
81	tmean_10	Temp prom 10	Mean temp 10	Bioclimatic
82	tmean_11	Temp prom 11	Mean temp 11	Bioclimatic
83	tmean_12	Temp prom 12	Mean temp 12	Bioclimatic
84	tmean_2	Temp prom 2	Mean temp 2	Bioclimatic
85	tmean_3	Temp prom 3	Mean temp 3	Bioclimatic
86	tmean_4	Temp prom 4	Mean temp 4	Bioclimatic
87	tmean_5	Temp prom 5	Mean temp 5	Bioclimatic
88	tmean_6	Temp prom 6	Mean temp 6	Bioclimatic
89	tmean_7	Temp prom 7	Mean temp 7	Bioclimatic
90	tmean_8	Temp prom 8	Mean temp 8	Bioclimatic
91	tmean_9	Temp prom 9	Mean temp 9	Bioclimatic

Continued

ID	CODE	VARIABLE (Local mode)	VARIABLE (On server mode)	COMPONENT
92	tmin_1	Temp min 1	Min temp 1	Bioclimatic
93	tmin_10	Temp min 10	Min temp 10	Bioclimatic
94	tmin_11	Temp min 11	Min temp 11	Bioclimatic
95	tmin_12	Temp min 12	Min temp 12	Bioclimatic
96	tmin_2	Temp min 2	Min temp 2	Bioclimatic
97	tmin_3	Temp min 3	Min temp 3	Bioclimatic
98	tmin_4	Temp min 4	Min temp 4	Bioclimatic
99	tmin_5	Temp min 5	Min temp 5	Bioclimatic
100	tmin_6	Temp min 6	Min temp 6	Bioclimatic
101	tmin_7	Temp min 7	Min temp 7	Bioclimatic
102	tmin_8	Temp min 8	Min temp 8	Bioclimatic
103	tmin_9	Temp min 9	Min temp 9	Bioclimatic
104	POINT_X	Longitud	Longitude	Geophysical
105	POINT_Y	Latitud	Latitude	Geophysical
106	vapr_1	Presion de vapor 1	Vap press 1	Bioclimatic
107	vapr_2	Presion de vapor 2	Vap press 2	Bioclimatic
108	vapr_3	Presion de vapor 3	Vap press 3	Bioclimatic
109	vapr_4	Presion de vapor 4	Vap press 4	Bioclimatic
110	vapr_5	Presion de vapor 5	Vap press 5	Bioclimatic
111	vapr_6	Presion de vapor 6	Vap press 6	Bioclimatic
112	vapr_7	Presion de vapor 7	Vap press 7	Bioclimatic
113	vapr_8	Presion de vapor 8	Vap press 8	Bioclimatic
114	vapr_9	Presion de vapor 9	Vap press 9	Bioclimatic
115	vapr_10	Presion de vapor 10	Vap press 10	Bioclimatic
116	vapr_11	Presion de vapor 11	Vap press 11	Bioclimatic
117	vapr_12	Presion de vapor 12	Vap press 12	Bioclimatic
118	vapr_annual	Presion de vapor anual	Vap press annual	Bioclimatic
119	wind_1	Velocidad viento 1	Wind speed 1	Geophysical
120	wind_2	Velocidad viento 2	Wind speed 2	Geophysical
121	wind_3	Velocidad viento 3	Wind speed 3	Geophysical
122	wind_4	Velocidad viento 4	Wind speed 4	Geophysical

## Continued

ID	CODE	VARIABLE (Local mode)	VARIABLE (On server mode)	COMPONENT
123	wind_5	Velocidad viento 5	Wind speed 5	Geophysical
124	wind_6	Velocidad viento 6	Wind speed 6	Geophysical
125	wind_7	Velocidad viento 7	Wind speed 7	Geophysical
126	wind_8	Velocidad viento 8	Wind speed 8	Geophysical
127	wind_9	Velocidad viento 9	Wind speed 9	Geophysical
128	wind_10	Velocidad viento 10	Wind speed 10	Geophysical
129	wind_11	Velocidad viento 11	Wind speed 11	Geophysical
130	wind_12	Velocidad viento 12	Wind speed 12	Geophysical
131	wind_annual	Velocidad viento anual	Wind speed annual	Geophysical
132	srad_1	Radiacion solar 1	Solar radiat 1	Geophysical
133	srad_2	Radiacion solar 2	Solar radiat 2	Geophysical
134	srad_3	Radiacion solar 3	Solar radiat 3	Geophysical
135	srad_4	Radiacion solar 4	Solar radiat 4	Geophysical
136	srad_5	Radiacion solar 5	Solar radiat 5	Geophysical
137	srad_6	Radiacion solar 6	Solar radiat 6	Geophysical
138	srad_7	Radiacion solar 7	Solar radiat 7	Geophysical
139	srad_8	Radiacion solar 8	Solar radiat 8	Geophysical
140	srad_9	Radiacion solar 9	Solar radiat 9	Geophysical
141	srad_10	Radiacion solar 10	Solar radiat 10	Geophysical
142	srad_11	Radiacion solar 11	Solar radiat 11	Geophysical
143	srad_12	Radiacion solar 12	Solar radiat 12	Geophysical
144	srad_annual	Radiacion solar anual	Solar radiat annual	Geophysical
145	t_awc1	Avail soil water cap h1 top	Avail soil water cap h1 top	Edaphic
146	s_awc1	Avail soil water cap h1 sub	Avail soil water cap h1 sub	Edaphic
147	t_awc2	Avail soil water cap h2 top	Avail soil water cap h2 top	Edaphic
148	s_awc2	Avail soil water cap h2 sub	Avail soil water cap h2 sub	Edaphic
149	t_awc3	Avail soil water cap h3 top	Avail soil water cap h3 top	Edaphic
150	s_awc3	Avail soil water cap h3 sub	Avail soil water cap h3 sub	Edaphic
151	t_awcts	Sat water cont top	Sat water cont top	Edaphic
152	s_awcts	Sat water cont sub	Sat water cont sub	Edaphic
153	depth_rock	Depth to bedrock	Depth to bedrock	Edaphic

## Continued

ID	CODE	VARIABLE ( <i>Local mode</i> )	VARIABLE ( <i>On server mode</i> )	COMPONENT
154	r_horizon	R horizon	R horizon	Edaphic
155	t_bulk_dens	Bulk density top	Bulk density top	Edaphic
156	s_bulk_dens	Bulk density sub	Bulk density sub	Edaphic
157	t_cecsol	Cation exchange cap top	Cation exchange cap top	Edaphic
158	s_cecsol	Cation exchange cap sub	Cation exchange cap sub	Edaphic
159	t_clay_cont	Clay content top	Clay content top	Edaphic
160	s_clay_cont	Clay content sub	Clay content sub	Edaphic
161	t_coarse_frag	Coarse fragments top	Coarse fragments top	Edaphic
162	s_coarse_frag	Coarse fragments sub	Coarse fragments sub	Edaphic
163	t_oc_dens	Organic carbon dens top	Organic carbon dens top	Edaphic
164	s_oc_dens	Organic carbon dens sub	Organic carbon dens sub	Edaphic
165	t_oc_stock	Organic carbon stock top	Organic carbon stock top	Edaphic
166	s_oc_stock	Organic carbon stock sub	Organic carbon stock sub	Edaphic
167	t_oc_cont	Organic carbon content top	Organic carbon content top	Edaphic
168	s_oc_cont	Organic carbon content sub	Organic carbon content sub	Edaphic
169	t_ph_hox	Soil pH H2O top	Soil pH H2O top	Edaphic
170	s_ph_hox	Soil pH H2O sub	Soil pH H2O sub	Edaphic
171	t_ph_kcl	Soil pH KCl top	Soil pH KCl top	Edaphic
172	s_ph_kcl	Soil pH KCl sub	Soil pH KCl sub	Edaphic
173	sodicity	Sodic soil grade	Sodic soil grade	Edaphic
174	t_silt_cont	Silt content top	Silt content top	Edaphic
175	s_silt_cont	Silt content sub	Silt content sub	Edaphic
176	t_sand_cont	Sand content top	Sand content top	Edaphic
177	s_sand_cont	Sand content sub	Sand content sub	Edaphic
178	t_soilwater_cap	Avail soil water cap top	Avail soil water cap top	Edaphic
179	s_soilwater_cap	Avail soil water cap sub	Avail soil water cap sub	Edaphic
178	t_soilwater_cap	Avail soil water cap top	Disp agua punto marchitez top	Edáfica
179	s_soilwater_cap	Avail soil water cap sub	Disp agua punto marchitez sub	Edáfica

## 19.2. Description of the variables available in CAPFITOGEN3 (ID is a common field in table 19.1).

ID	DESCRIPTION	SOURCE
1	Elevation. Meters above the sea level	<a href="http://worldclim.org">http://worldclim.org</a>
2	Aspect (degree) of the land. 0 and 359 degrees correspond to north.	NA
3	Annual Mean Temperature	<a href="http://worldclim.org">http://worldclim.org</a>
4	Mean Temperature of Warmest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
5	Mean Temperature of Coldest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
6	Annual Precipitation	<a href="http://worldclim.org">http://worldclim.org</a>
7	Precipitation of Wettest Month	<a href="http://worldclim.org">http://worldclim.org</a>
8	Precipitation of Driest Month	<a href="http://worldclim.org">http://worldclim.org</a>
9	Precipitation Seasonality (Coefficient of Variation)	<a href="http://worldclim.org">http://worldclim.org</a>
10	Precipitation of Wettest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
11	Precipitation of Driest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
12	Precipitation of Warmest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
13	Precipitation of Coldest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
14	Mean Diurnal Range (Mean of monthly (max temp - min temp))	<a href="http://worldclim.org">http://worldclim.org</a>
15	Isothermality (BIO2/BIO7) (* 100)	<a href="http://worldclim.org">http://worldclim.org</a>
16	Temperature Seasonality (standard deviation * 100)	<a href="http://worldclim.org">http://worldclim.org</a>
17	Max Temperature of Warmest Month	<a href="http://worldclim.org">http://worldclim.org</a>
18	Min Temperature of Coldest Month	<a href="http://worldclim.org">http://worldclim.org</a>
19	Temperature Annual Range (BIO5-BIO6)	<a href="http://worldclim.org">http://worldclim.org</a>
20	Mean Temperature of Wettest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
21	Mean Temperature of Driest Quarter	<a href="http://worldclim.org">http://worldclim.org</a>
22	Eastness	NA
23	Northness	NA
24	Precipitation January	<a href="http://worldclim.org">http://worldclim.org</a>
25	Precipitation October	<a href="http://worldclim.org">http://worldclim.org</a>
26	Precipitation November	<a href="http://worldclim.org">http://worldclim.org</a>
27	Precipitation December	<a href="http://worldclim.org">http://worldclim.org</a>
28	Precipitation February	<a href="http://worldclim.org">http://worldclim.org</a>
29	Precipitation March	<a href="http://worldclim.org">http://worldclim.org</a>



## Continued

ID	DESCRIPTION	SOURCE
30	Precipitation April	<a href="http://worldclim.org">http://worldclim.org</a>
31	Precipitation May	<a href="http://worldclim.org">http://worldclim.org</a>
32	Precipitation June	<a href="http://worldclim.org">http://worldclim.org</a>
33	Precipitation July	<a href="http://worldclim.org">http://worldclim.org</a>
34	Precipitation August	<a href="http://worldclim.org">http://worldclim.org</a>
35	Precipitation September	<a href="http://worldclim.org">http://worldclim.org</a>
36	Reference depth of the soil unit	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
37	Subsoil Base Saturation	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
38	Subsoil Calcium Carbonate	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
39	Subsoil Gypsum	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
40	Subsoil CEC (clay)	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
41	Subsoil CEC (soil)	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
42	Subsoil Clay Fraction	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
43	Subsoil Salinity	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
44	Subsoil Sodicity	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
45	Subsoil Gravel Content	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
46	Subsoil Organic Carbon	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
47	Subsoil pH (H <sub>2</sub> O)	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
48	Subsoil Reference Bulk Density	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
49	Subsoil Sand Fraction	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
50	Subsoil Silt Fraction	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
51	Subsoil total exchangeable bases	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
52	Slope (in degrees) of the land surface	NA
53	Topsoil Base Saturation	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
54	Topsoil Calcium Carbonate	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
55	Topsoil Gypsum	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
56	Topsoil CEC (clay)	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
57	Topsoil CEC (soil)	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
58	Topsoil Clay Fraction	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
59	Topsoil Salinity	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
60	Topsoil Sodicity	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )

**Continued**

ID	DESCRIPTION	SOURCE
61	Topsoil Gravel Content	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
62	Topsoil Organic Carbon	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
63	Topsoil pH (H2O)	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
64	Topsoil Reference Bulk Density	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
65	Topsoil Sand Fraction	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
66	Topsoil Silt Fraction	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
67	Topsoil total exchangeable bases	HWSD ( <a href="http://t.ly/dbry">http://t.ly/dbry</a> )
68	Maximum temperature January	<a href="http://worldclim.org">http://worldclim.org</a>
69	Maximum temperature October	<a href="http://worldclim.org">http://worldclim.org</a>
70	Maximum temperature November	<a href="http://worldclim.org">http://worldclim.org</a>
71	Maximum temperature December	<a href="http://worldclim.org">http://worldclim.org</a>
72	Maximum temperature February	<a href="http://worldclim.org">http://worldclim.org</a>
73	Maximum temperature March	<a href="http://worldclim.org">http://worldclim.org</a>
74	Maximum temperature April	<a href="http://worldclim.org">http://worldclim.org</a>
75	Maximum temperature May	<a href="http://worldclim.org">http://worldclim.org</a>
76	Maximum temperature June	<a href="http://worldclim.org">http://worldclim.org</a>
77	Maximum temperature July	<a href="http://worldclim.org">http://worldclim.org</a>
78	Maximum temperature August	<a href="http://worldclim.org">http://worldclim.org</a>
79	Maximum temperature September	<a href="http://worldclim.org">http://worldclim.org</a>
80	Mean temperature January	<a href="http://worldclim.org">http://worldclim.org</a>
81	Mean temperature October	<a href="http://worldclim.org">http://worldclim.org</a>
82	Mean temperature November	<a href="http://worldclim.org">http://worldclim.org</a>
83	Mean temperature December	<a href="http://worldclim.org">http://worldclim.org</a>
84	Mean temperature February	<a href="http://worldclim.org">http://worldclim.org</a>
85	Mean temperature March	<a href="http://worldclim.org">http://worldclim.org</a>
86	Mean temperature April	<a href="http://worldclim.org">http://worldclim.org</a>
87	Mean temperature May	<a href="http://worldclim.org">http://worldclim.org</a>
88	Mean temperature June	<a href="http://worldclim.org">http://worldclim.org</a>
89	Mean temperature July	<a href="http://worldclim.org">http://worldclim.org</a>
90	Mean temperature August	<a href="http://worldclim.org">http://worldclim.org</a>
91	Mean temperature September	<a href="http://worldclim.org">http://worldclim.org</a>

## Continued

ID	DESCRIPTION	SOURCE
92	Minimum temperature January	<a href="http://worldclim.org">http://worldclim.org</a>
93	Minimum temperature October	<a href="http://worldclim.org">http://worldclim.org</a>
94	Minimum temperature November	<a href="http://worldclim.org">http://worldclim.org</a>
95	Minimum temperature December	<a href="http://worldclim.org">http://worldclim.org</a>
96	Minimum temperature February	<a href="http://worldclim.org">http://worldclim.org</a>
97	Minimum temperature March	<a href="http://worldclim.org">http://worldclim.org</a>
98	Minimum temperature April	<a href="http://worldclim.org">http://worldclim.org</a>
99	Minimum temperature May	<a href="http://worldclim.org">http://worldclim.org</a>
100	Minimum temperature June	<a href="http://worldclim.org">http://worldclim.org</a>
101	Minimum temperature July	<a href="http://worldclim.org">http://worldclim.org</a>
102	Minimum temperature August	<a href="http://worldclim.org">http://worldclim.org</a>
103	Minimum temperature September	<a href="http://worldclim.org">http://worldclim.org</a>
104	Longitude for the cell centroid	NA
105	Latitude for the cell centroid	NA
106	Water vapor pressure January	<a href="http://worldclim.org">http://worldclim.org</a>
107	Water vapor pressure February	<a href="http://worldclim.org">http://worldclim.org</a>
108	Water vapor pressure March	<a href="http://worldclim.org">http://worldclim.org</a>
109	Water vapor pressure April	<a href="http://worldclim.org">http://worldclim.org</a>
110	Water vapor pressure May	<a href="http://worldclim.org">http://worldclim.org</a>
111	Water vapor pressure June	<a href="http://worldclim.org">http://worldclim.org</a>
112	Water vapor pressure July	<a href="http://worldclim.org">http://worldclim.org</a>
113	Water vapor pressure August	<a href="http://worldclim.org">http://worldclim.org</a>
114	Water vapor pressure September	<a href="http://worldclim.org">http://worldclim.org</a>
115	Water vapor pressure October	<a href="http://worldclim.org">http://worldclim.org</a>
116	Water vapor pressure November	<a href="http://worldclim.org">http://worldclim.org</a>
117	Water vapor pressure December	<a href="http://worldclim.org">http://worldclim.org</a>
118	Water vapor pressure Annual	<a href="http://worldclim.org">http://worldclim.org</a>
119	Wind speed January	<a href="http://worldclim.org">http://worldclim.org</a>
120	Wind speed February	<a href="http://worldclim.org">http://worldclim.org</a>
121	Wind speed March	<a href="http://worldclim.org">http://worldclim.org</a>
122	Wind speed April	<a href="http://worldclim.org">http://worldclim.org</a>

## Continued

ID	DESCRIPTION	SOURCE
123	Wind speed May	<a href="http://worldclim.org">http://worldclim.org</a>
124	Wind speed June	<a href="http://worldclim.org">http://worldclim.org</a>
125	Wind speed July	<a href="http://worldclim.org">http://worldclim.org</a>
126	Wind speed August	<a href="http://worldclim.org">http://worldclim.org</a>
127	Wind speed September	<a href="http://worldclim.org">http://worldclim.org</a>
128	Wind speed October	<a href="http://worldclim.org">http://worldclim.org</a>
129	Wind speed November	<a href="http://worldclim.org">http://worldclim.org</a>
130	Wind speed December	<a href="http://worldclim.org">http://worldclim.org</a>
131	Wind speed Annual	<a href="http://worldclim.org">http://worldclim.org</a>
132	Solar radiation January	<a href="http://worldclim.org">http://worldclim.org</a>
133	Solar radiation February	<a href="http://worldclim.org">http://worldclim.org</a>
134	Solar radiation March	<a href="http://worldclim.org">http://worldclim.org</a>
135	Solar radiation April	<a href="http://worldclim.org">http://worldclim.org</a>
136	Solar radiation May	<a href="http://worldclim.org">http://worldclim.org</a>
137	Solar radiation June	<a href="http://worldclim.org">http://worldclim.org</a>
138	Solar radiation July	<a href="http://worldclim.org">http://worldclim.org</a>
139	Solar radiation August	<a href="http://worldclim.org">http://worldclim.org</a>
140	Solar radiation September	<a href="http://worldclim.org">http://worldclim.org</a>
141	Solar radiation October	<a href="http://worldclim.org">http://worldclim.org</a>
142	Solar radiation November	<a href="http://worldclim.org">http://worldclim.org</a>
143	Solar radiation December	<a href="http://worldclim.org">http://worldclim.org</a>
144	Solar radiation Annual	<a href="http://worldclim.org">http://worldclim.org</a>
145	Available soil water capacity (volumetric fraction) for h1 - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
146	Available soil water capacity (volumetric fraction) for h1 - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
147	Available soil water capacity (volumetric fraction) for h2 - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
148	Available soil water capacity (volumetric fraction) for h2 - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
149	Available soil water capacity (volumetric fraction) for h3 - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
150	Available soil water capacity (volumetric fraction) for h3 - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
151	Saturated water content (volumetric fraction) for tS - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
152	Saturated water content (volumetric fraction) for tS - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
153	Depth to bedrock (R horizon) up to 200 cm	<a href="https://soilgrids.org">https://soilgrids.org</a>

## Continued

ID	DESCRIPTION	SOURCE
154	Probability of occurrence of R horizon	<a href="https://soilgrids.org">https://soilgrids.org</a>
155	Bulk density (fine earth) in kg / cubic-meter - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
156	Bulk density (fine earth) in kg / cubic-meter - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
157	Cation exchange capacity of soil in cmolc/kg - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
158	Cation exchange capacity of soil in cmolc/kg - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
159	Clay content (0-2 micro meter) mass fraction in % - topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
160	Clay content (0-2 micro meter) mass fraction in % - subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
161	Coarse fragments volumetric in % topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
162	Coarse fragments volumetric in % subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
163	Soil organic carbon density in kg per cubic-m topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
164	Soil organic carbon density in kg per cubic-m subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
165	Soil organic carbon stock in tons per ha topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
166	Soil organic carbon stock in tons per ha subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
167	Soil organic carbon content (fine earth fraction) in g per kg topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
168	Soil organic carbon content (fine earth fraction) in g per kg subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
169	Soil pH in H2O topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
170	Soil pH in H2O subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
171	Soil pH in KCl topsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
172	Soil pH in KCl subsoil	<a href="https://soilgrids.org">https://soilgrids.org</a>
173	Sodic soil grade	<a href="https://soilgrids.org">https://soilgrids.org</a>
174	Silt content (2-50 micro meter) mass fraction in %	<a href="https://soilgrids.org">https://soilgrids.org</a>
175	Silt content (2-50 micro meter) mass fraction in %	<a href="https://soilgrids.org">https://soilgrids.org</a>
176	Sand content (50-2000 micro meter) mass fraction in %	<a href="https://soilgrids.org">https://soilgrids.org</a>
177	Sand content (50-2000 micro meter) mass fraction in %	<a href="https://soilgrids.org">https://soilgrids.org</a>
178	Available soil water capacity (volumetric fraction) until wilting point	<a href="https://soilgrids.org">https://soilgrids.org</a>
179	Available soil water capacity (volumetric fraction) until wilting point	<a href="https://soilgrids.org">https://soilgrids.org</a>

### 19.3. Explanation of extra columns in the 'tabla\_de\_analisisGEOQUAL.txt/xls' table.

Variable	Explanation
globlandc	Extracted value from GLC 2000 (Global Land Cover 2000).
DISTOLAND	Anillo de distancia a la tierra dentro del cual las coordenadas se encuentran. (0 = tierra, 1 = 1 km, 10 = 10 km, etc.).
SUITQUAL	SUITQUAL parameter (0 to 20 values).
ID_0	Extracted values from GADM identifying the area of the country.
ISO	Extracted values from GADM compared with ORIGCTY.
NAME_0	Extracted values from GADM for the country's full name.
ID_1	Extracted values from GADM identifying the area at the NAME_1 level.
NAME_1	Extracted values from GADM compared with ADM1.
VARNAME_1	Extracted values from GADM for alternative names to NAME_1.
ENGTYPE_1	Extracted values from GADM defining the type of administration represented by NAME_1.
ID_2	Extracted values from GADM identifying the area at the NAME_2 level.
NAME_2	Extracted values from GADM compared with ADM2.
VARNAME_2	Extracted values from GADM for alternative names to NAME_2.
ENGTYPE_2	Extracted values from GADM defining the type of administration represented by NAME_2.
ID_3	Extracted values from GADM identifying the area at the NAME_3 level.
NAME_3	Extracted values from GADM compared with ADM3.
VARNAME_3	Extracted values from GADM for alternative names to NAME_3.
ENGTYPE_3	Extracted values from GADM defining the type of administration represented by NAME_3.
ID_4	Extracted values from GADM identifying the area at the NAME_4 level.
NAME_4	Extracted values from GADM compared with ADM4.
VARNAME_4	Value extracted from GADM for alternative names to NAME_4.
ENGTYPE4	Value extracted from GADM defining the type of administration represented by NAME_4.
NIVELMAX	Depending on the country, this is the lowest administrative level included in GADM.
LOCALQUAL	LOCALQUAL parameter (values 0 to 20).
COORQUAL	COORQUAL parameter (values 0 to 20).
intertemp	COORQUAL intertemp sub-parameter
errors	COORQUAL errors sub-parameter
precis	COORQUAL precis sub-parameter
georable	COORQUAL georable sub-parameter

**Continued**

TOTALQUAL	TOTALQUAL parameter (values from 0 to 40 or 0 to 60, depending on whether LOCALQUAL is included or not).
TOTALQUAL100	TOTALQUAL100 parameter (values from 0 to 100).



---

<http://www.capfitogen.net/en/>

---