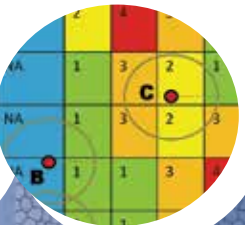
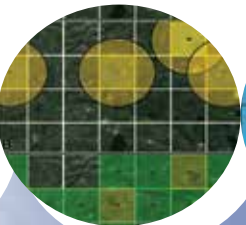
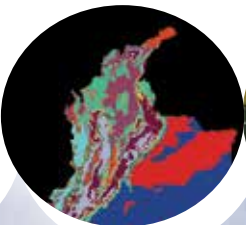
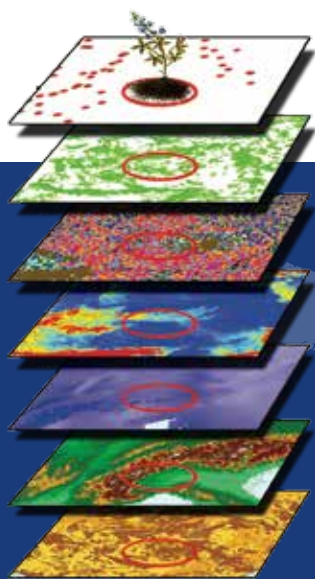
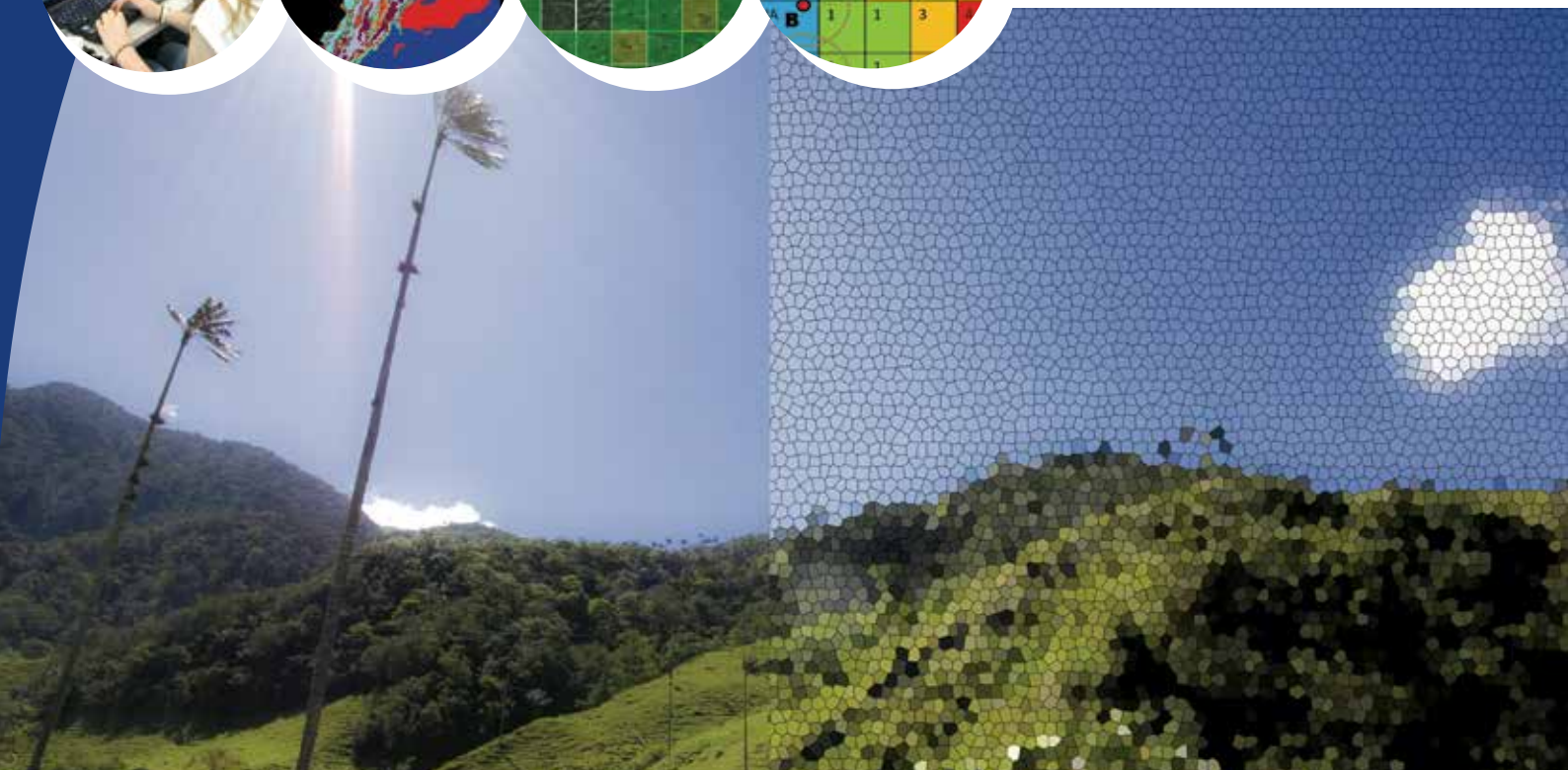




The International Treaty  
ON PLANT GENETIC RESOURCES FOR FOOD AND AGRICULTURE



*Tools*



# CAPFITOGEN

Programme to Strengthen  
National Plant Genetic  
Resource Capacities  
in Latin America

Version 1.2



*Tools*

# CAPFITOGEN

Programme to Strengthen  
National Plant Genetic  
Resource Capacities  
in Latin America

**Versión 1.2**

**Author of the tools:**

Mauricio Parra Quijano

*Consultant*

*International Treaty on Plant Genetic Resources  
for Food and Agriculture, (ITPGRFA)*

*FAO*

**Authors of the accompanying manual:**

Mauricio Parra Quijano

Elena Torres Lamas, *Universidad Politécnica de Madrid (Spain)*

José María Iriondo Alegría, *Universidad Rey Juan Carlos (Spain)*

Francisco López, *ITPGRFA, FAO*

The terms used in this information product and the form in which the data contained appear, do not imply in any way, that the United Nations Food and Agriculture Organisation has any judgment concerning the legal status or level of development of countries, territories, cities or zones, or their authorities, or with respect to the delimitation of their frontiers or territorial limits. The mention of any specific companies or products manufactured, whether or not they are patented, does not mean that the FAO endorses these or is recommending others of a similar nature which are not mentioned.

The opinions expressed in this information product are those of its authors and do not necessarily reflect the views or policies of the FAO.

ISBN

© FAO 2014

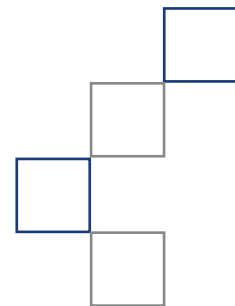
The FAO promotes the use, reproduction and dissemination of the material contained in this information product. Unless indicated to the contrary, the material may be copied, printed and downloaded for the purposes of private study, research and teaching, or for use in products or services for no commercial purpose, as-and-when the FAO is acknowledged as author and owner of the copyright, and that this in no way means that the FAO endorses the views, products or services of the users.

Any queries concerning translation and adaptation rights as well as the resale and other rights of commercial use should be addressed to [www.fao.org/contact-us/licence-request](http://www.fao.org/contact-us/licence-request) or a [copyright@fao.org](mailto:copyright@fao.org).

FAO information products are available on its website ([www.fao.org/publications](http://www.fao.org/publications)) and may be acquired by sending an e-mail to [publications-sales@fao.org](mailto:publications-sales@fao.org).

Electronic products

FAO may not be held liable for errors or deficiencies in the database, software or accompanying documentation, nor for program maintenance and upgrading, nor for any damage that may arise from these. Similarly, the FAO shall not be held responsible for updating the data and is not liable in any way for any errors or omissions in the data provided. Nonetheless, the FAO requests users to report any errors or deficiencies that they may find.



---

## Contents

<b>1.</b>	Programme to Strengthen National Plant Genetic Resource Capacities in Latin America	5
<b>2.</b>	CAPFITOGEN Tools: features and installation	9
<b>3.</b>	GEOQUAL Tool	23
<b>4.</b>	ELCmapas Tool	35
<b>5.</b>	ECOGEO Tool	49
<b>6.</b>	Representa Tool	59
<b>7.</b>	DIVmapas Tool	71
<b>8.</b>	ColNucleo Tool	95
<b>9.</b>	FIGS_R Tool	105
<b>10.</b>	Frequent errors	119
<b>11.</b>	Acknowledgments	127
<b>12.</b>	Annexes	133





# 1. Programme to Strengthen National Plant Genetic Resource Capacities in Latin America

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 on implementing 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). The success of the events was a testament to the effectiveness of this kind of workshops in contributing to the implementation of the ITPGRFA objectives within the GRULAC community. The coordination of the workshops between the organizations involved was a decisive factor in achieving the objectives set, particularly given Spain's commitment to the ITPGRFA. The organizations involved were the Secretariat of the ITPGRFA, the Spanish International Cooperation and Development agency (*Agencia Española para la Cooperación Internacional y el Desarrollo - AECID*), the Spanish Ministry of Environment, Rural and Marine areas (*Ministerio de Medio Ambiente y Medio Rural y Marino*) and the Spanish Plant Genetic Resources Center (Centro Nacional de Recursos Fitogenéticos) of the Spanish Institute for Research and Agrarian Technology (Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria - CRF-INIA). The success of the events was also due to the close relationship between the National Plant Genetic Resources Conservation programs and the National Agricultural Research Institutes (Institutos Nacionales de Investigación Agrícola - INIAs) in Spain and the GRULAC countries.

The positive experiences from earlier workshops and the importance of achieving some key ITPGRFA objectives in Latin America and the Caribbean, particularly those explained in articles 5, 6, 7, 8 and 13.2 c, were strong incentives to continue with these activities. At the same time, it became clear that there was a need for the workshops to be developed in greater depth with more technical content by setting up a technology transfer program where the workshops would be a key element of a broader-based action strategy.

Taking this precedent and the region's necessities as a point of departure, the Programme to Strengthen National Plant Genetic Resource Capacities in Latin America - CAPFITOGEN - was launched. This program is focused on the development of appropriate technologies for countries which are extremely agrobiodiverse but have limited economic resources. Its function is to develop and transfer technology and provide the appropriate training for technical personnel from those Latin American countries signatories to the Treaty.





The warm reception given to the tools and methodologies developed under the auspices of the CAPFITOGEN program in 2013, has meant that some countries targeted by the program have organized national workshops on their own initiative, financed by the most interested parties. At the same time, there has been interest from other countries and regions not initially targeted by the program, which have been asking for tools and transfer and training activities.

Thus the program CAPFITOGEN is primed to function not only as a generator and facilitator of appropriate technology, but also as a model of transfer in itself. One of its most innovative aspects is the way in which it seeks to involve people who have developed scientific methodologies. They are invited to develop the tools provided by the program based on their methodologies and to carry out the technical training and transfer activities themselves. This model means that the program beneficiaries are guaranteed direct access to the scientists and developers in order to answer queries or discuss cases. At the same time, the scientists themselves benefit directly from the experiences and issues tackled by the technical experts from the national programs, an outcome with a positive impact on future investigations with a more focused application and better suited to meet real needs.







---

## 2. CAPFITOGEN tools: features and installation

### 2.1. Origin

Cultivated plants today were once wild plants whose array of genetic weaponry helped them withstand and adapt to the challenges brought by a constantly-changing environment, such as plagues, disease, a grazing herbivore, drought, etc. The process of domestication transformed these wild plants into the grains, legumes, fruits and vegetables that we know today; a selection of products tailored to meet the needs and tastes of human beings. However, achieving these products involved a lengthy and intense selection process. As a result, the genetic basis of these domesticated species is rather small if we compare it with that of their wild ancestors.

The evolution of cultivated species has left behind it a range of products among which there are also differences. For example, the modern varieties and hybrids which began to be developed as from the 1960s are material with a high production potential but an overly-narrow genetic base when compared with varieties from the early XX century.

Thus domestication, a largely selective process, has left many genetic configurations by the wayside in its search for the “best” variety. This process led to the loss of many valuable genes that could had provided solutions to future problems we may encounter with our modern and productive varieties, which are yet vulnerable and homogenous. Fortunately, not everything has disappeared from the fields, and farmers around the world continue to preserve heirloom varieties inherited from their ancestors. These are unlikely to be commercially viable but have a tremendous significance for their cultural values, eating and food habits and even religious traditions.

Aware of the progressive loss of their genetic heritage, many nations began to rescue and preserve these varieties and the wild plants related to them, storing them in germplasm collections (seeds, plant tissue, propagules) as from the 1950s.

#### 2.1.1 Conservation to increase knowledge and use

Germplasm collections are essentially different from a museum’s collection as what is preserved is intended to be used at a later date. The main users tend to work in the area of crop breeding and seek out features of interest from specimens in germplasm collections with a view to transferring these to modern varieties. However, they can only employ germplasm efficiently and effectively on the basis

of the knowledge derived from it. Obtaining this knowledge is carried out through specific processes of characterization and evaluation which require a substantial investment in economic and logistics terms by those entities in charge of conserving agrobiodiversity. At a national level, these activities are carried out as part of national programs; in Latin America, they are usually the premise of the national institutes of agricultural research (Institutos Nacionales de Investigación Agraria - INIAs).

### 2.1.2 Appropriate techniques for scenarios with limited resources

Methods to collect, preserve, and characterize agrobiodiversity using scientific standards have generally been developed in regions and centers where there are few, if any, restraints on investment, infrastructure, or staff qualifications. This has meant that developing countries are either directly unable to apply these methodologies or, if they do so, they are nonetheless unable to include all the preserved germplasm. This situation is in stark contrast to the fact that it is precisely in developing countries where the greatest concentrations of agricultural plant genetic wealth are to be found.

This scenario has prompted some research groups in the world to direct their efforts towards the exploration of less costly and complex methodologies which are far better adapted to the context of national programs in developing countries. Such alternative methodologies include making use of environmental information from collection sites (ecogeographic), for example, to evaluate the genetic variability of the germplasm, or calculate with a higher success margin the probability of locating genes of interest. Similarly, another option is the use of geographic information systems as a means of obtaining and making use of these ecogeographic data. Given that most ecogeographic information and computer programs for performing analysis are free, the size of the investment is reduced to a computer with an ordinary business-type configuration and the training required for the personnel involved. These are, then, methods which are compatible with a scenario of limited resources, a recurring condition of national programs in developing countries. After demonstrating their effectiveness in case studies published in international scientific journals, the new generation of methodologies were chosen to be adapted and transformed into simplified tools and practices before being finally transferred to national programs in the regions selected.

## 2.2 Characteristics

Ecogeographic applications and geographic information systems (GIS) encompass a range of processes directed at the conservation and efficient use of plant genetic resources. These, as demonstrated in the articles published in various scientific media, involve the use of numerous software programs, many of which are neither

easily accessible nor free, to achieve complex methodological analyses. Thus, the main challenge lies in developing practical tools that enable these advances to be applied by technical personnel with no training in either managing and programming statistics or GIS software.

The solution lies in the use of a computer program that offers the following features:

- (a) Includes all the statistical analyses required.
- (b) Includes all GIS functions, both for the management of georeferenced data and their subsequent analysis.
- (c) Is able to manage databases and the products of statistical analysis and GIS.
- (d) Enables all the results of the analysis to be saved in conventional computer formats.
- (e) Be freely-available and widely-distributed.

Currently, the “R” software programming environment (R. Cran, 2012), widely used in the scientific community, offers all the features necessary to develop the CAPFITOGEN tools. The “R” software environment is very powerful for analysis but its main drawback is that it requires specific technical knowledge in order to program its functions properly. The CAPFITOGEN program has overcome this drawback by involving the research team responsible for the most of the original methodologies to develop the program’s tools, particularly the R programming. The R program seeks to integrate all the functions and the analysis required in one single program in such a way that it operates with predefined data formats, through which any user data can be entered into the R routines to achieve specific results.

The last challenge in the popularization of the use of these tools was to simplify the way in which commands and data are entered. This has been solved thanks to a simple interface based on Java programming and html with the use of Tomcat virtual servers . This solution is offered by the RWUI application (<http://sysbio.mrc-bsu.cam.ac.uk/Rwui>), which has undergone some minor modifications in order to be used for the development of the CAPFITOGEN tools.

## 2.3 Installation and execution

Version 1.1 of the CAPFITOGEN tools required a complex installation process including the manual installation of R, the Java execution and Tomcat virtual server environments, as well as the modification of certain Windows environment variables. The old procedure required the user to manually install the programs and changes in the environment variables, as well as take into account the type of operating system (32 bit or 64 bit) of the Windows version.

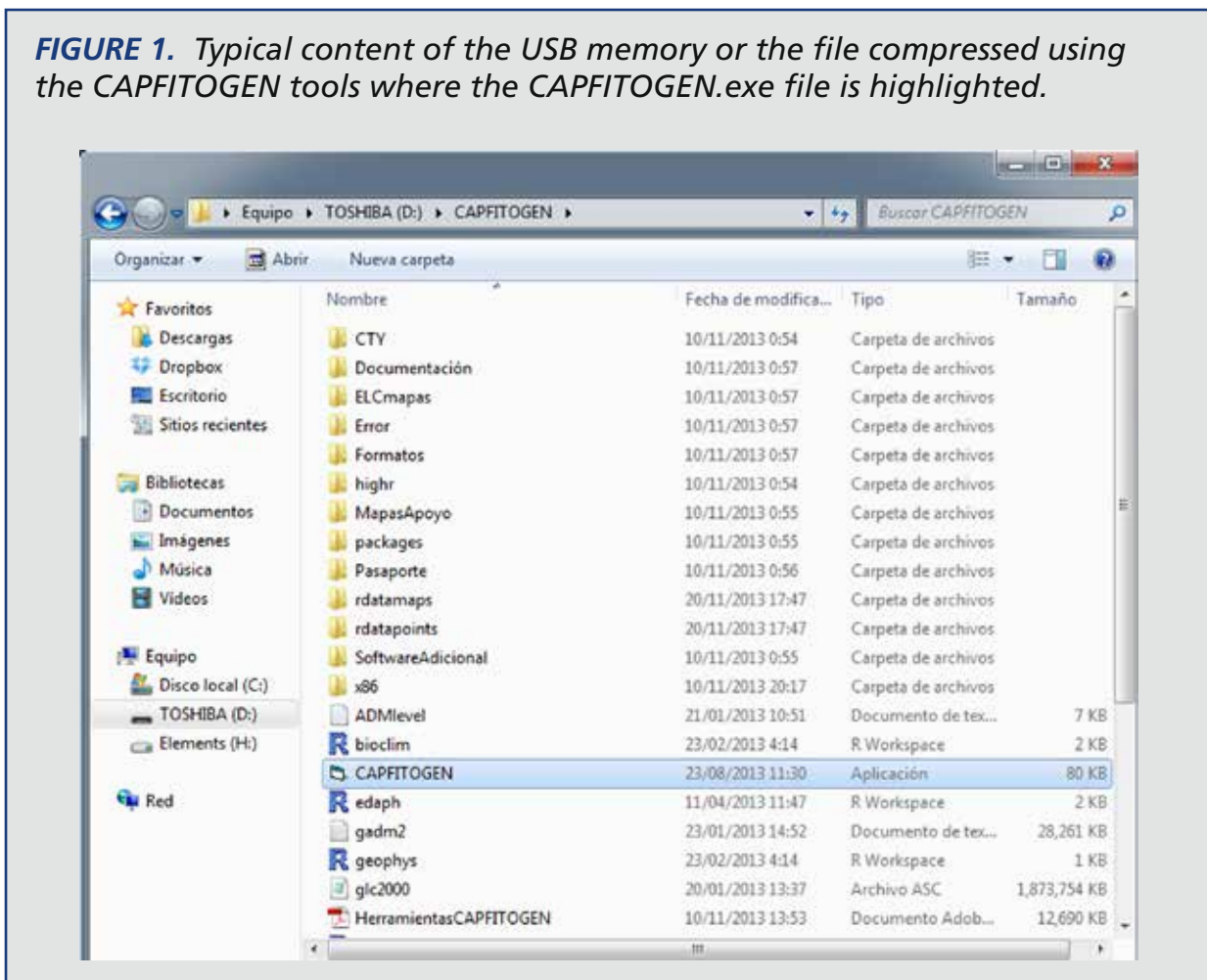


From version 1.2 onwards, the CAPFITOGEN tool set comes with an installer which performs all the steps required to install the program and modify the environment variables. This installer is also responsible for unzipping and installing all the R packages required to perform the analyses.

The tools package has been developed to be ready to use. The steps required to install and execute the CAPFITOGEN tools are the following:

- (a) If you have downloaded the tools from the web site created for this purpose (see <http://www.agrobiodiversidad.org/blog/?p=1039>), you should have an .alz extension file (compressed) with the main body of tools and other .alz files with the ecogeographic information for each country or region contemplated as a work area. Extract all files or folders with those files, and then assemble the Tools installer, placing the folder with the name of the country or region in the "rdatamaps" folder in the set of files and folders in the main body of the tool, which is included in the "CAPFITOGEN" folder. Copy the folder, preferably onto an external drive dedicated specifically to tools, or directly

**FIGURE 1.** Typical content of the USB memory or the file compressed using the CAPFITOGEN tools where the CAPFITOGEN.exe file is highlighted.



onto the hard disk of the PC, frequently corresponding to the C:\ drive. The systems used to download material from internet and set up the tools may change in subsequent versions.

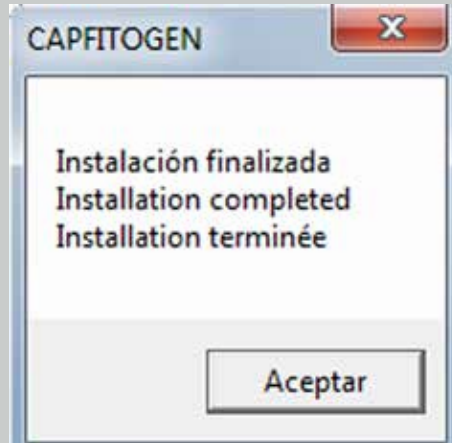
- (b) If you have has accessed the tools using a USB flash drive provided by the CAPFITOGEN program, simply connect it to one of your PC's USB ports. The contents of the USB drive include a folder called "CAPFITOGEN". Do not move the contents of this folder. This folder contains the structure of folders and files (Fig. (1) necessary to operate the tools.
- (c) in the set of files you will find one single executable file (with the file extension .exe), usually called "CAPFITOGEN.exe". Double click on this file and a window will immediately open up like the one shown in Fig. 2.
- (d) Click on the "Install" button. The installer will show the progress of the installation of the different programs in blue. When it has finished, a gray-colored window will appear, indicating the completion of the process (Fig. 3). Click on the "OK" button in this window. It is not necessary to restart the computer.
- (e) Immediately after the installation has finished, the initial window (which showed the "Install" button) will now show the "Execute" button. Click on this link.
- (f) At this point, a black-colored window will appear and a series of white code numbers will be loaded (see Fig. 4) This code means the program to enable the tomcat virtual server is being loaded. The instructions are complete when the following line of text appears, indicating the milliseconds (ms) taken to activate the virtual server:

INFO: Server startup in xxxx ms

**FIGURE 2.** Window showing the installation of the CAPFITOGEN tools.



**FIGURE 3.** Window indicating that the installation is complete.



- (g) A few seconds after beginning the previous process, the default internet browser program will open automatically (for example: Internet Explorer, Mozilla Firefox or Google Chrome) as shown in Fig. 5a (language selection panel) and Fig. 5b (tool selection panel). The following instruction will appear in the browser address bar:

`http://localhost:8080/start/`

**FIGURE 4.** A black window where the virtual server is being loaded. This window should not be closed until the tools have finished being used.

```

Tomcat
-----
at org.apache.catalina.startup.HostConfig.deployDirectory(HostConfig.jav
a:1114)
at org.apache.catalina.startup.HostConfig$DeployDirectory.run(HostConfig
.java:1673)
at java.util.concurrent.Executors$RunnableAdapter.call(Unknown Source)
at java.util.concurrent.FutureTask$Sync.innerRun(Unknown Source)
at java.util.concurrent.FutureTask.run(Unknown Source)
at java.util.concurrent.ThreadPoolExecutor.runWorker(Unknown Source)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(Unknown Source)
at java.lang.Thread.run(Unknown Source)

ene 17, 2014 9:02:10 AM org.apache.catalina.startup.HostConfig deployDirectory
INFO: Despliegue del directorio D:\CAPFITOGEN\x86\tomcat_x86\apache-tomcat-7-0-3
5_x86\webapps\ROOT de la aplicaci3n web
ene 17, 2014 9:02:10 AM org.apache.catalina.startup.HostConfig deployDirectory
INFO: Despliegue del directorio D:\CAPFITOGEN\x86\tomcat_x86\apache-tomcat-7-0-3
5_x86\webapps\start de la aplicaci3n web
ene 17, 2014 9:02:10 AM org.apache.coyote.AbstractProtocol start
INFO: Starting ProtocolHandler ["http-bio-8080"]
ene 17, 2014 9:02:10 AM org.apache.coyote.AbstractProtocol start
INFO: Starting ProtocolHandler ["ajp-bio-8009"]
ene 17, 2014 9:02:10 AM org.apache.catalina.startup.Catalina start
INFO: Server startup in 2377 ms
  
```

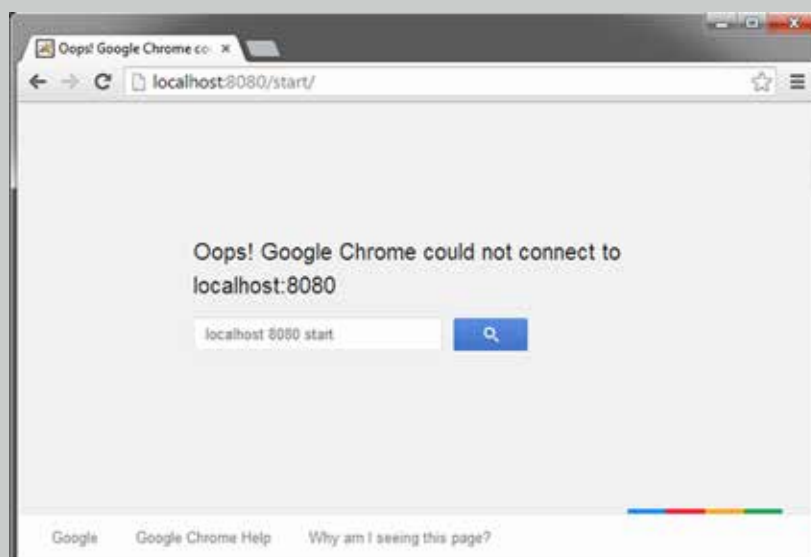
Occasionally, the process described above in point “f” begins before the process described in point “e” has finished. This means that the virtual server is not ready when summoned by the browser, which prompts an error message to appear in the browser (see fig. 6) Process “f” occurs before “e”, occurs frequently the first time that the tools are run or when the computer’s configuration is out of date. This problem is solved simply by closing and opening the browser and typing the following in the address bar:

`http://localhost:8080/start/`

**FIGURE 5.** The browser opens automatically, showing: a) the language selection panel and b) the tool selection panel.



**FIGURE 6.** Example of an error occurring when the browser attempts to open the virtual server address before it completed the loading process.





(h) A start panel will appear in the browser (Fig. 5), showing the list of tools available on the left-hand side and on the languages available for each tool on the right. Clicking on the green icons opens the form for the tool and language selected in the browser. In Fig. 7 the form for the GEOQUAL tool is shown as an example. This form displays a small bar with links as described in Fig. 8.

To exit the application once you have finished using the tool, simply close your browser, close the black background window and close the window of Fig. 2 by clicking on the "Exit" button. It is possible that, after closing all the windows, Windows may ask you if the application was correctly installed. Please answer in the affirmative.

**FIGURE 7. GEOQUAL tool form. 1. Languages available and links concerning legal issues. 2. Space indicating the tool currently in use. 3. Links bar. 4. Brief description of the tool. 5. Link to the instruction manual. 6. Area where parameters may be introduced. 7. Button to start the scan.**

1 Nota legal Derechos de uso

2

3

4 This application allows the quality of geo-referencing at the data collection sites described in the passport data to be evaluated. This means evaluating both the locations description and its coordinates. The passport data must be in the same format as FAO/Bioversity 2012 (which is very similar to FAO IPGRI 2001). You may use an Excel format which is to be found in the folder TablaPasaporteModelo. The assessment is carried out by determining three quality parameters (LOCALQUAL, COORQUAL and SUITQUAL), which are combined to give a single quality value called TOTQUAL100. TOTQUAL100 quality values range from 0 to 100, where 0 corresponds to cases with neither coordinates nor a location description (zero quality) and 100 is the optimum value. Author of the tool: Mauricio Parra Quijano, mauricio.parra@fao.org. International Treaty on Plant Genetic Resources for Food and Agriculture, 2013.

5 Manual

6 Path where the CAPFITOGEN tools are located. Note: use / instead of \. For example F:\, C:\CAPFITOGEN, D:\mytools\CAPFITOGEN, etc.

6 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 Passport folder, which is part of the set of folders that make up the GEOQUAL tool.

6 You may choose whether to use high or low resolution maps to determine whether the coordinates for a collection site are in the sea and, if so, at what distance. High resolution may slow the process down in very large databases (over 15,000 entries with coordinates).

6 Please indicate if you wish to use the LOCALQUAL parameter to evaluate georeferencing quality. LOCALQUAL is a parameter used to compare a locabems description with the location drawn by GIS. If your data includes no location description or the description is only contained in the COLLSITE field, it is better to use this option.

6 Insert the path to the folder where the results of the analysis will be saved. Note: use / instead of \. For example C:\CAPFITOGEN, D:\mytools\CAPFITOGEN, etc.

7 Analyze

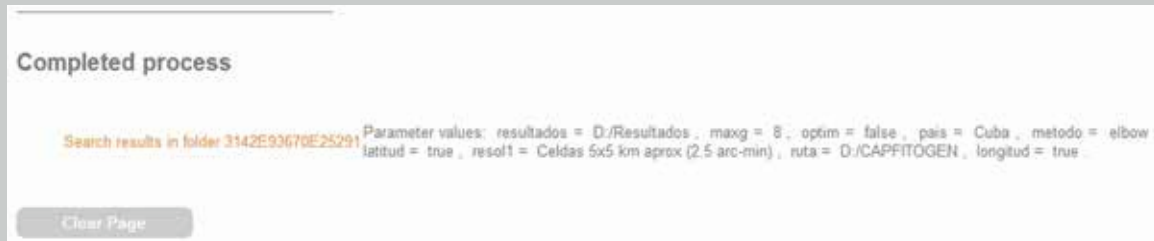
When the installation is carried out, an identification file is added to the drive and path where the files and folders of the CAPFITOGEN tools are kept. This file enables the computer to recognize that the installation has already been carried out. In this way, each time you need to use the tools, just click on the CAPFITOGEN.exe file for the installation window to open, displaying the "Execute" button. From this step onwards, the entire process from point "d" as indicated above is repeated. If for any reason the letter of the drive or path where the executable file is located changes or the ID file is deleted accidentally, when you double click on CAPFITOGEN.exe the "Install" button will be displayed again. It is advisable at this point to reinstall the program and tools. If R and Java programs are reinstalled, there is no danger of any problems arising in the execution of the tools.

**FIGURE 8.** Links bar (corresponds to part 3 of Fig.6). The buttons have the following functions: 1. Return to the form for the tool selected. 2. See the CAPFITOGEN tools instruction manual 3. Contact and support information. 4. Return to the page where the tools may be selected (landing page).




(i) While the analysis is running, Windows will display the standby icon as the mouse pointer of the mouse. Once the analysis has been successfully completed, the page containing the form will move to the top of the screen, showing the header with the ITPGRFA logo. When you go to the bottom of the page, a "Process finished" message will appear (Fig.9), indicating that the previously saved results are now available. If you wish to change any parameter and rerun the analysis, you can do so using the same form by changing the parameter in question and clicking on the "Analyze" button. You can also delete the entire contents of the form with the "Clear Page" button.

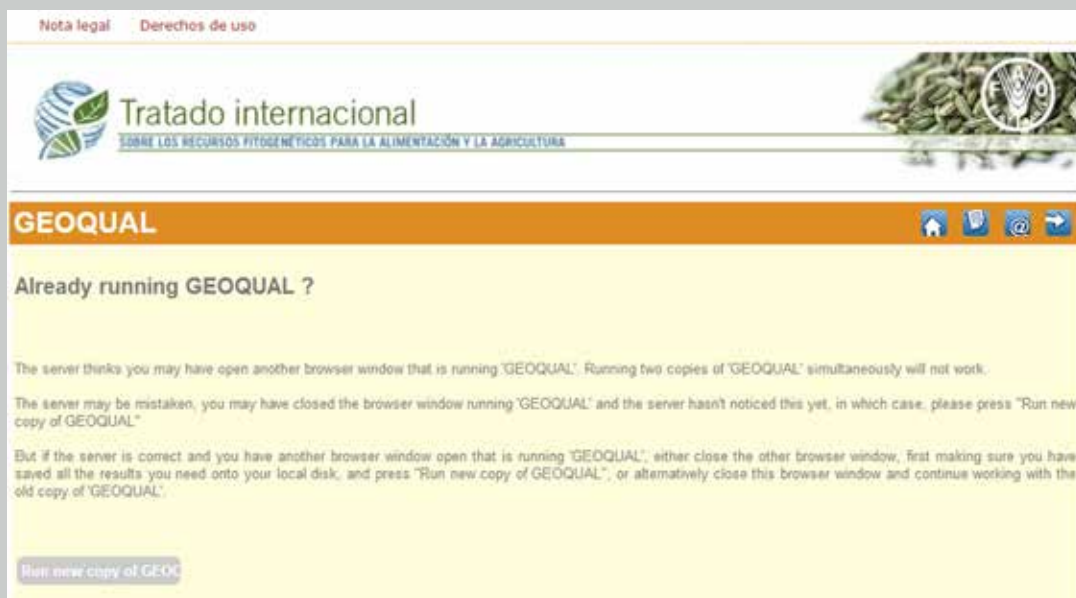
**FIGURE 9.** View of the bottom of the form page once the analysis has been completed correctly and the results are available.



The results of the analysis are to be found in the folder defined by the user in the “results” parameter. Maps, both in raster format (mesh of cells with .grd and .gri file extensions), such as the maps generated by the ELCmapas tool and those in vector format (usually “shapefiles”), for example, point maps, may be opened with DIVA-GIS software since they are fully compatible. Results tables are generated in text format separated by tabs, which can be opened with programs such as Microsoft Excel or the Open Office spreadsheets.

- (j) It is possible that when exiting a tool to go to the start panel (using the button  in Fig.8 and subsequently returning to the same tool, or if the address is simply reloaded in the browser, a window will appear like that shown in Fig. 10 with instructions in English. This window asks the user if they wish to reload a tool that has previously been loaded. No problems will occur if the tool is reloaded from this window. All that is required to continue is to click on the “Run new copy of” button and the name of the tool.
- (k) When an error occurs in the execution of the tool, either due to a faulty installation, or a parameter that was incorrectly entered, or because of defects in the tables providing passport or characterization information, the page with the form shows the header (as it does when the process has finished successfully), but the time it take is considerably less than a successful process and an error message appears at the bottom of the page, as shown in Fig. 11. It is possible to detect the source of the error through the error message. Thus, when the message (which is directly generated by the R program) appears, its contents should be compared with the list of messages in Chapter 10 “Frequent errors”.

**FIGURE 10.** A window that indicates that you are attempting to reload a tool which was open before. To continue, simply click on the “Run new copy of” button and the name of the tool.



**FIGURE 11.** A view of the bottom of the page of the form when an error has occurred.



- (l) It is possible that not all the R packages required to carry out the planned analysis during the installation are properly loaded due to the different settings of some personal computers. This will produce a recurrent error when the tools are executed, which will show up in red as in Fig. 11 but with the following text:

**An error occurred: Error in the library (package name): there is no package called 'name of package' Calls: source -> withVisible -> eval -> eval -> library**



This error, which is detailed in Chapter 10 “Frequent errors”, may be repeated for all the packages needed, making the process of correction proposed in Chapter 10 too extensive. In this case, the user should delete the “library” folder to be found in the C:\rwin route, then copy the “library” folder from the CAPFITOGEN\Error route and paste it into the C:\rwin route. This action will completely eliminate the problems arising from the absence of those packages necessary for R.

(m) In some instances, due to a faulty installation process or the incorrect location of the “CAPFITOGEN” folder, it may be that the interface is unable to activate R and send the information necessary to perform the analysis. This problem will become evident when, after clicking on the button “Analyze” and having filled in the parameters correctly, the execution time is abnormally short and error messages do not appear in red. Instead, the message of “Process finished” appears and no new file appears in the folder where the results are usually saved. In this case make sure of two aspects:

1. The files and folders making up the CAPFITOGEN tools are stored in a folder and are not directly to be found in the root directory. In other words, the path to capfitogen.exe should be X:\CAPFITOGEN\capfitogen.exe and not X:\capfitogen.exe, where X is the letter of the disk drive. If the tools need to be reinstalled, ensure the condition detailed above is met.

2. If the CAPFITOGEN folder is not in the root directory (X:\CAPFITOGEN) but in a path in other folders, something which is in itself not to be recommended, this path should not contain any spaces between words. For example, if the folder is stored in the X:\Mis Documentos\CAPFITOGEN path, the error will show up because of the space between the words “Mis” and “Documentos”.

If, in spite of making sure that the two conditions above are met, the problem persists, please contact the program’s technical support team.

n) When the process executed after clicking on the “Analyze” button takes too long (over 15 or 20 minutes) it is possible that the page displaying the tool form will change its appearance and display a Java error message, such as the one shown in Fig. 12. The page change and error message do not indicate that the process has terminated abruptly nor that results will not be generated. This error is more about the visualization of the interface than the R process itself. Therefore, the next step is to check if the files expected appear as results in the designated folder.

**FIGURE 12.** Java error notice that appears when the process takes too long.



```
← → C localhost:8080/ELCmapas/enterdata.do
Estado HTTP 500 - java.lang.IllegalStateException: setAttribute: La Sesión ya ha sido invalidada

Type Informe de Excepción
mensaje java.lang.IllegalStateException: setAttribute: La Sesión ya ha sido invalidada
descripción El servidor encontró un error interno que hizo que no pudiera rellenar este requerimiento.
excepción
javax.servlet.ServletException: java.lang.IllegalStateException: setAttribute: La Sesión ya ha sido invalidada
    org.apache.struts.action.RequestProcessor.processException(RequestProcessor.java:535)
    org.apache.struts.action.RequestProcessor.processActionPerform(RequestProcessor.java:433)
    org.apache.struts.action.RequestProcessor.process(RequestProcessor.java:236)
    org.apache.struts.action.ActionServlet.process(ActionServlet.java:1196)
    org.apache.struts.action.ActionServlet.doPost(ActionServlet.java:432)
    javax.servlet.http.HttpServlet.service(HttpServlet.java:647)
    javax.servlet.http.HttpServlet.service(HttpServlet.java:728)

causa raíz
java.lang.IllegalStateException: setAttribute: La Sesión ya ha sido invalidada
    org.apache.catalina.session.StandardSession.setAttribute(StandardSession.java:1437)
    org.apache.catalina.session.StandardSession.setAttribute(StandardSession.java:1402)
    org.apache.catalina.session.StandardSessionFacade.setAttribute(StandardSessionFacade.java:156)
    org.apache.struts.action.SynchroAction.execute(SynchroAction.java:170)
    org.apache.struts.action.RequestProcessor.processActionPerform(RequestProcessor.java:431)
    org.apache.struts.action.RequestProcessor.process(RequestProcessor.java:236)
    org.apache.struts.action.ActionServlet.process(ActionServlet.java:1196)
    org.apache.struts.action.ActionServlet.doPost(ActionServlet.java:432)
    javax.servlet.http.HttpServlet.service(HttpServlet.java:647)
    javax.servlet.http.HttpServlet.service(HttpServlet.java:728)

nota: la traza completa de la causa de este error se encuentra en los archivos de diario de Apache Tomcat/7.0.35.

Apache Tomcat/7.0.35
```





---

## 3. GEOQUAL Tool

### 3.1. 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 make an assessment of 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 directly 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 the data beyond 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 from ecology and spatial analysis to the patterns of the distribution of species. There are several studies that point out that quality is a critical issue in methodologies such as the modeling 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 prior to any analysis which 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.



### 3.2. History of the GEOQUAL tool

The methodology which 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 characterized ecogeographically. 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, <http://www.sierfe.es>).

SIERFE is a system which enables the selection of germplasm on the basis of the environmental characterization of a collection site through an internet portal. With the development of GEOQUAL and its incorporation into SIERFE, a quality estimator allows SIERFE users (seekers of germplasm, such as breeders, scientists, or farmers) to define their requirements in terms of the quality of geo-referencing when selecting germplasm by ecogeographic variables. This represents a major advance in the development of information systems and germplasm selection. Over 45,000 accessions in the Spanish inventory have been ecogeographically characterized and each one given a quality rating value on a scale from 0 to 100.

GEOQUAL was then 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 have become 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 CAPFITOGEN program was approved and the tools to be developed were decided, GEOQUAL became a priority. This was about addressing the challenge of creating a tool capable of evaluating the quality of geo-referencing data, a simple tool which already had all the necessary information preloaded,

and which did not require a great knowledge of geographic information systems (GIS) in order to apply it. 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, and 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 which offers a range of adaptability factors which are appropriate for the conditions and needs of various national programs for the conservation of plant genetic resources.

### 3.3. Features of GEOQUAL

The 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. In addition, 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 (see Annex 9.1). However, if the data were in the 2001 FAO-IPGRI format, GEOQUAL would also be able operate after migrating the 2001 formatting information to 2012 without having to add information for new fields included by the 2012 version. However, it would have to consider including the four ADM descriptors.

The 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 (folder "DescriptoresPasaporteModelo", file "Tabla pasaporte modelo FAO\_Bioversity 2012 modificada.xls"), where the color green is used to identify descriptors which are essential for GEOQUAL and yellow for those which, although not essential, are nonetheless important. Non-designated fields are not taken into account by GEOQUAL but their position in the table should nonetheless be maintained (as in the case of those listed) so that GEOQUAL is able to find the variables it needs to analyze 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.

### 3.3.1 Description of GEOQUAL's base parameters

#### 3.3.1.1 COORQUAL parameter

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

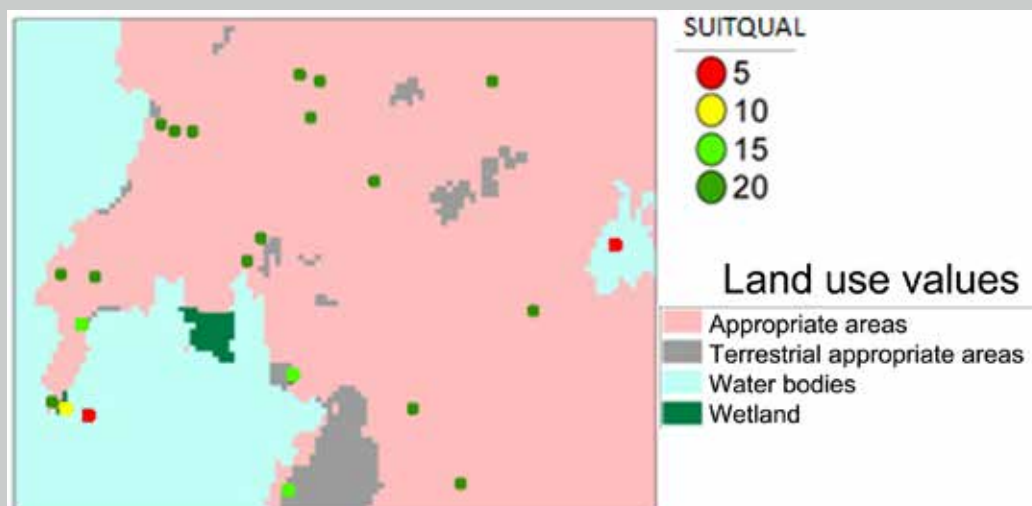
- (a) ERRORS: If the coordinates in decimal or sexagesimal format contain values out of the references of the WGS84 latlong coordinates system. Uses the descriptors LATITUDE, LONGITUDE, DECLATITUDE and DECLONGITUDE.
- (b) PRECIS: This applies to coordinates in decimal format that comply with the coding 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. Use the following descriptors:
- (c) GEORBLE: The descriptor evaluates the possibility of obtaining 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 which occurred after 2000, it is highly likely that GPS was used, which would increase the quality of the coordinates.
- (e) GEOREFMETH: This 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 taken into account 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 the maximum quality. The combined values of each sub-parameter generate the COORQUAL parameter in a range of zero to twenty.

#### 3.3.1.2 SUITQUAL parameter

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) use. This is an older and freely accessible global coverage map which provides details on the use of land with a 1 km resolution. 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 13.** An example of how to obtain *SUITQUAL* values according to the interpretation of land use values.



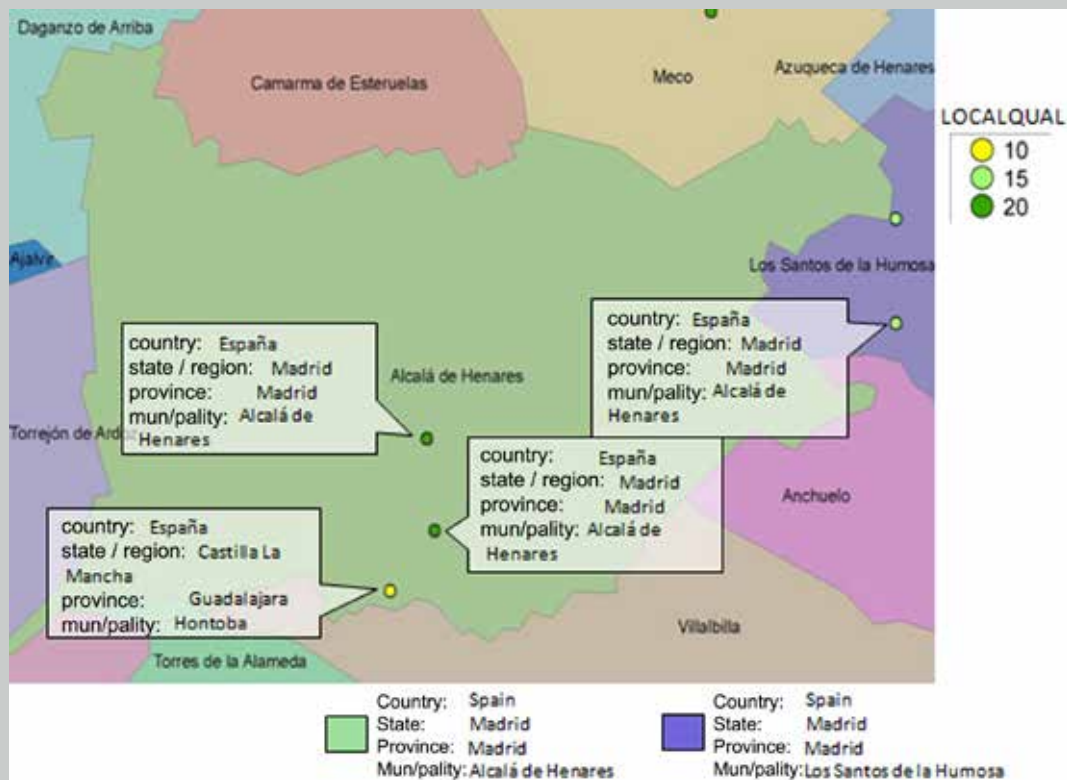
### 3.3.1.3 LOCALQUAL parameter

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 database. 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 performs "Check Coordinates" (checking coordinates) included in DIVA-GIS where the comparison is absolute (the terms must match character for 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 is able to identify concordances despite typographical errors or differences created by using alphabetical characters from certain languages which 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



**FIGURE 14.** 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.



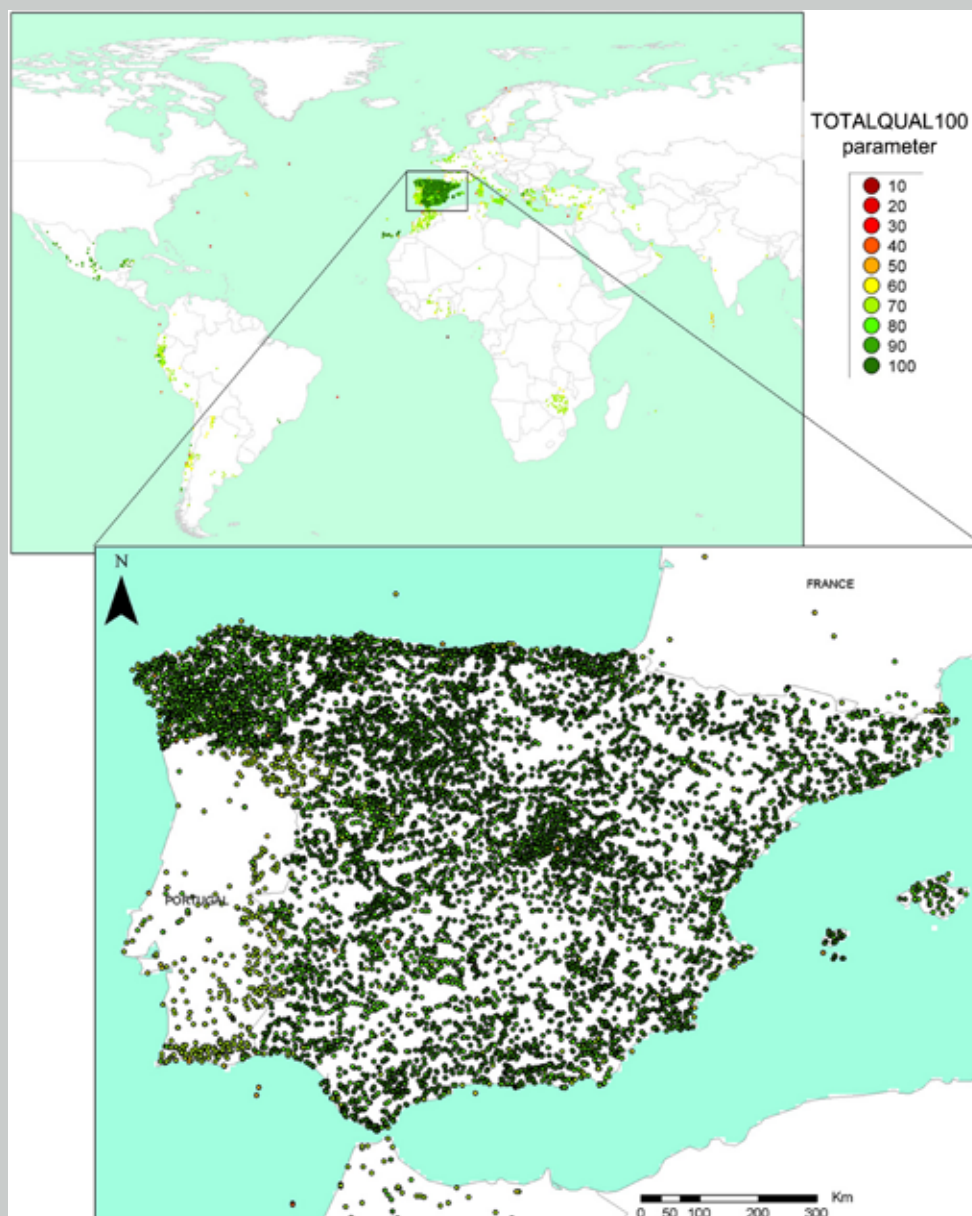
registering germplasm in their passport databases.

Lastly, LOCALQUAL takes into account 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.

### 3.3.2 Description of the TOTALQUAL and TOTALQUAL100 parameters

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 is able to work with values from 0 to 60. However, to make it easier to interpret and analyze the results generated by GEOQUAL, the TOTALQUAL100 parameter is also calculated. This is a transformation of TOTALQUAL to a range of values from 0 to 100, where 0 is zero quality (including

**FIGURE 15.** Results of the application of GEOQUAL to the Spanish National Inventory of Plant Genetic Resources. The values reached by TOTALQUAL100 are displayed.



the lack of coordinates) and 100 represents a theoretically optimum quality.

### 3.3.3 Determination of quality thresholds

Since it was first put to use, GEOQUAL has been designed to be used 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 geo-referenced 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 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, in order 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.

### 3.4. Using the GEOQUAL tool

Once the CAPFITOGEN tools have been installed and the GEOQUAL tool selected, it will be necessary to define a set of parameters to ensure the R program runs correctly.

Following the definition of all the parameters and routes required by GEOQUAL, the analytical process will begin after clicking on the "Analyze" button. After a time which 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 (3.4.1.6 parameter).

#### 3.4.1 Initial parameters defined by the user

##### 3.4.1.1 Parameter: *ruta*

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

##### 3.4.1.2 Parameter: *pasaporte*

Explanation: 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 "passport" folder which is part of the set of folders making up the CAPFITOGEN directory.

#### 3.4.1.3 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)

#### 3.4.1.4 Parameter: local

Explanation: Specify whether you wish to use the LOCALQUAL parameter 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 COLLSITE field, this option is UNSUITABLE.

#### 3.4.1.5 Parameter: resultados

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.

## 3.5. GEOQUAL results

In the path and folder created for "resultados" (parameter 3.4.1.5) there should be three tables and a map of vector-type points (shapefile).

### 3.5.1 Tables

The tables generated by GEOQUAL are in tab-delimited text format and can be opened in programs such as Excel, OpenOffice, or R.

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

3.5.1.2 "tabla\_de\_analisisGEOQUAL.txt": This table also contains all columns of the passports table which were originally introduced for the purposes of analysis, 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 which correspond to extractions, interpretations or sub-parameters and which are considered necessary to calculate the values of the GEOQUAL parameters. The list of additional variables included in this table and their explanation are found in Annex 12.4.

### 3.5.2 Maps

3.5.2.1 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 colors 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.

3.5.2.2. 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.

### 3.6. 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. Multi-crop Passport descriptors 2001. List developed by FAO and IPGRI.

FAO, BIOVERSITY. 2012. FAO/Bioversity Multi-crop Passport descriptors V.2. Available at [http://www.bioversityinternational.org/index.php?id=19&user\\_bioversitypublications\\_pi1%5BshowUid%5D=6901](http://www.bioversityinternational.org/index.php?id=19&user_bioversitypublications_pi1%5BshowUid%5D=6901)

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 geo-reference the world’s biodiversity data. BMC Bioinformatics. 2009 Nov 10; 10 Suppl 14:S3. DOI: 10.1186/1471-2105-10-S14-S3.

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

Soberon, j., Peterson, T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Phil. Trans- R Soc. LOND. B. 359, 689-698.



## 4. ELCmapas Tool

### 4.1. Applications of the Ecogeographic Land Characterization (ELC) mapping tool

ELC mapping tools provide information on various different environmental scenarios in order to assess plant adaptation processes in a given territory. They are also useful for conservation and the sustainable 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 usually 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 favorable 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 in studies of climate change, for example. 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, in addition to 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 on the basis of 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 which explicitly refers to adaptation. One early reference is an ecogeographic map drawn up in 1997 to help create core collections (Tohmé et al., 1995), although other different criteria were also taken into account after the accessions 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 software with a high capacity for analysis now retails at a discount. Furthermore,

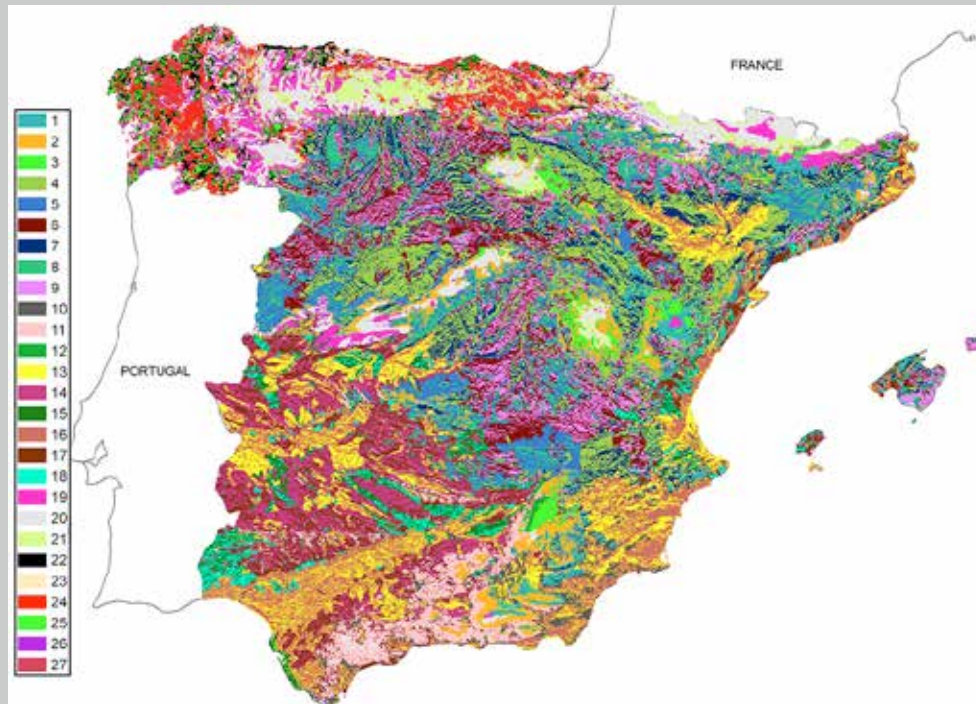


access to the internet has increased markedly in developing countries. This progress has impacted on the development of maps representing different adaptation scenarios for crop wild relatives and was responsible for the generation of the first ecogeographic land characterization map for Spain in 2005 (Parra Quijano et al., 2008). This is a general map which 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, in order 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 on the basis of other sources of ecogeographic information, although the methodology used was similar to the one produced in 2005. The researchers were keen to both ensure that the map was able to portray adaptive scenarios as faithfully as possible, and establish whether it could be used to perform an evaluation. They evaluated the performance of the new map with eight species (four leguminous varieties and four grasses), two of which were crop wild relatives (CWR), while the other six were local varieties. Adaptive values were assigned to the distribution and "seed weight" variable used as a variable phenotypic indicator. The results were compared with 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 taking into account aspects related to the abiotic adaptation of plants (CORINE land cover map, land use map, see <http://www.eea.europa.eu/data-and-maps/explore-interactive-maps/corine-landcover-2006>). 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, see <http://www.eea.europa.eu/data-and-maps/figures/dmeer-digital-map-of-european-ecological-regions>).

Overall, the ELC map performed better with the leguminous species than with the grasses, although the exception was *Zea mays* which elicited an acceptable result. As expected, the map displayed adaptive scenarios for both CWR, but also produced acceptable results for the local varieties, such as in *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 unadvisable when drawing conclusions from

**FIGURE 16.** ELC map of Peninsular Spain and the Balearic Islands for the *Lupinus* genus.



a large group of species, particularly if the map is not properly evaluated. It is also important to make a proper selection of the ecogeographic variables representing the three key abiotic aspects involved in plant development: bioclimatic, geophysical and edaphic aspects.

## 4.2. History of the ELCmapas tool

The ELCmapas tool covered in this manual represents the development of the concept Ecogeographical Land Characterization Maps 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 (e.g., Parra Quijano et al. 2011A, 2011 B, 2012 B and Thormann 2012).

The interest that this methodology prompted among various teams and research projects concerning the collection, conservation and use of plant genetic resources contrasted with a specific observation made repeatedly by potential users. The methodology described in this publication is complex because it mixes geographic

information systems (GIS) with multivariate analysis techniques. In addition, the original development implied the use of a commercial program to carry out statistical analysis. These issues were a major hindrance to the generation of ELC maps by researchers and technical experts.

### 4.3. Features of ELCmapas

The ELCmapas tool is a new option which uses R to develop ELC maps and also avoids the complications described above. This free software environment is able to compute large amounts of statistical data and has an impressive array of graphics resources able to integrate GIS with multivariate analysis. The tool can produce ELC maps without switching between different programs, 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 DIVA-GIS, Google Earth, or Microsoft Excel and thus these maps can be used as a component of other tools like Representa.

The ELCmapas tool uses two methods to determine the number of groups to use in the clustering analysis. These procedures are:

- a) A simple system that uses K - means as a clustering algorithm where the cut-off point is determined on the basis of the decrease in the sum of the intra-group squares (Ketchen and 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 method, also known as "elbow", as it can process large amounts of data without long delays, and is thus recommended for large countries.
- b) 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 the composition of the clusters to be checked (Kaufman and Rousseeuw, 1987; Rousseeuw, 1987). As this system consumes more computing resources, it takes considerably longer when applied to large data sets.

The methods used to determine the numbers of groups are not entirely objective, because the user decides the maximum number of groups allowed. Furthermore, the elbow method means that the percentage of decrease is subjective, even though it is based on the observation of the graphs of intra-group variance by the number of groups.

As ecogeographic information at a resolution of 1 km or even 5 km for an entire subcontinent such as Latin America is considerable, the ELCmapas tool is best used at country level, although the distribution of the species in question or

the distribution of germplasm collections may exceed national frontiers. With the new version (1.2) of ELCmapas, data with a lower resolution (10-20 km) is available. These resolutions can be used at continental or sub-continental level.

#### 4.4. How to select ecogeographic variables

The selection of ecogeographic variables needs to be established before using the ELCmapas tool. Any changes to a single variable of a single component (bioclimatic, geophysical or edaphic), or the addition or deletion of a variable, 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, it emerged that 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). Accordingly, a selection was made of the ecogeographic variables of each component with the greatest influence on the abiotic adaptation of the species and which thus determined their distribution.

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

- a) Bibliographic searches: It is easy to find references in technical and/or scientific publications about the environmental factors that influence, determine or limit the distribution of a species. Sometimes maps can be made on the basis of the correspondence between factors which use 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 in order to know which 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 more 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 on the basis of expert knowledge.



In this study, the map was used to determine the ideal location of genetic reserves for several Beta species in Europe.

After establishing the list of potential variables, each component (bioclimatic, geophysical and edaphic) must be analyzed to determine which ones are redundant. This involves performing an analysis of bivariate correlations or an analysis of collinearity. When there is a high correlation between two variables of the same component, one of them should be discarded. Furthermore, an analysis of the principal components (where all variables are quantitative) can help to define the relationships between variables and determine the final selection. No more than five variables should be used per component because 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 4.5.1.7 and 4.5.1.8) results in maps with larger areas and less cross-links. 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 the parameters bioclimv, geophysv and edaphv (parameters 4.5.1.5, 4.5.1.6 and 4.1.5.9). The complete list of variables, including the ELCmapas v1.2 tool, is given at the end of this document (Annexes 12.1, 12.2 and 12.3).

## 4.5. Using the ELCmapas Tool

Once the CAPFITOGEN tools have been installed and the ELCmapas tool selected, the user should specify a series of parameters.

### 4.5.1 Initial Parameters Defined by the User

#### 4.5.1.1 Parameter: *ruta*

Explanation: Path where the CAPFITOGEN tools have been copied or are to be found. Note: use / instead of \ when indicating the path of the folder. For example, L:/CAPFITOGEN, D:/CAPFITOGEN, etc.

#### 4.5.1.2 Parameter: *primvez*

Explanation: If this is the first time that the ELCmapas tool is used on this computer, click on this box.

#### 4.5.1.3 Parameter: *pais*

Explanation: Select the country for which you wish to build the ELC map. Multiple available countries will appear only if the tool includes information for all of them; otherwise, only one specific country will appear.

#### 4.5.1.4 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. See Annex 12.5 on the availability of resolutions in relation to the country or region selected.

#### 4.5.1.5 Parameter: *bioclimv*

Explanation: Select the bioclimatic variables (temperature, rainfall and indexes) that you wish to include in order to generate the ELC map. You can select a multiple of variables by holding down the Ctrl (control) key and adding more variables by clicking on them with the left-hand button on the mouse.

#### 4.5.1.6 Parameter: *geophysv*

Explanation: Select the geophysical variables (related to terrain and sunlight) that you wish to include in order to generate the ELC map. You can select multiple variables.

#### 4.5.1.7 Parameter: *latitud*

Explanation: Will you include latitude on your map? Note: by including latitude and longitude and excluding orientation, you will create more contiguous ecogeographic units and less cross-linking in the maps.

#### 4.5.1.8 Parameter: *longitud*

Explanation: Will you include longitude on your map? Note: by including latitude and longitude and excluding orientation, you will create more contiguous ecogeographic units and less cross-linking in the maps.

#### 4.5.1.9 Parameter: *edaphv*

Explanation: Select the edaphic variables (texture, depth, pH, etc.) that you wish to include in order to generate the ELC map. You can select multiple variables.

#### 4.5.1.10 Parameter: *optim*

Explanation: Please indicate if you require an optimization process. Optimization is only recommended for large countries (e.g., Argentina, Brazil, Mexico) using high (1x1 km). Otherwise, this may slow down the process. Do not use "optim" if you include latitude or longitude as variables to create the ELC map.

#### 4.5.1.11 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 five.

#### 4.5.1.12 Parameter: *metodo*

Explanation: Select one of the methods offered to generate the clusters in an objective manner. The elbow method is the simplest and fastest, while the medoids method is more sophisticated and requires more computing resources.

#### 4.5.1.13 Parameter: *resultados*

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.

Once all the parameters and paths required by ELCmapas have been defined, the analytical process will begin after clicking on the "Analyze" button. After some time, which may vary due to the specific resolution parameters, the type of analysis, the amount of processed data or the computer's hardware settings, the ELCmapas tool will produce results to be saved where indicated (parameter 4.5.1.13).

## 4.6. Results of ELCmapas

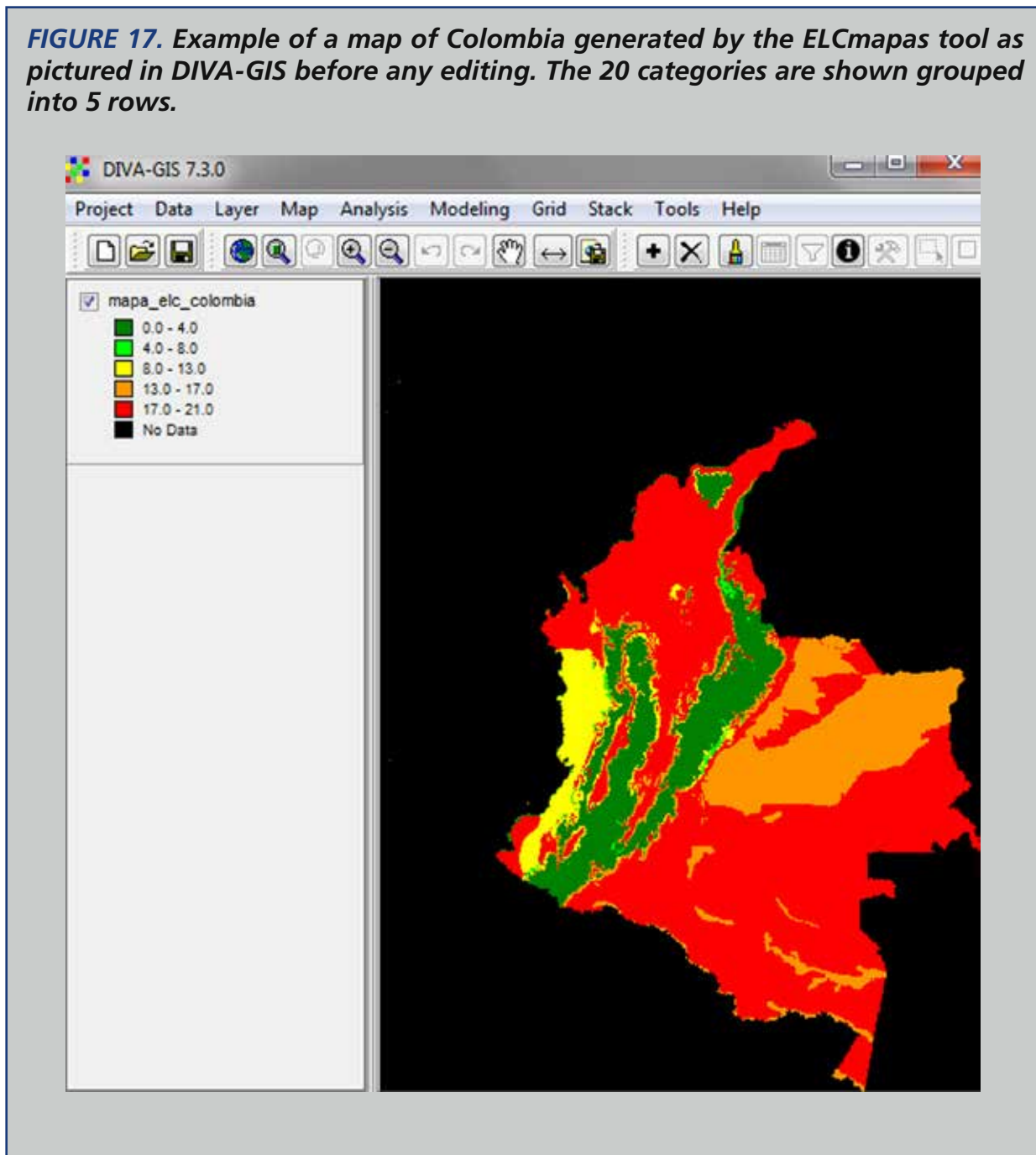
In the path and folder created for "resultados" (parameter 4.5.1.13) five maps and three tables are generated.

### 4.6.1 The Maps

These files correspond to the final ELC map of the country or the region determined in parameter 4.5.1.3 (mapa\_elc\_country.grd, and mapa\_elc\_country.gri, together with mapa\_elc\_country.png image) as well as the maps representing the categories resulting from the bioclimatic, geophysical and edaphic components (mapa\_bioclimatico\_country.grd, mapa\_bioclimatico\_country.gri, mapa\_geofisico\_country.grd, mapa\_geofisico\_country.gri, mapa\_edafico\_country.grd and mapa\_edafico\_country.gri). All of these maps can be opened in DIVA-GIS. Initially, DIVA-GIS opens the maps as shown in Fig. 17.

However, the display may be altered by double clicking on the gray panel on the left-hand side which represents this layer. By adding as many rows as there are categories in the map, and then applying a swatch of random colors, you can obtain a map such as that shown in Fig. 18. It helps to use widely-contrasting colors so that the categories (ecogeographic scenarios) present in the territory may be easily identified.

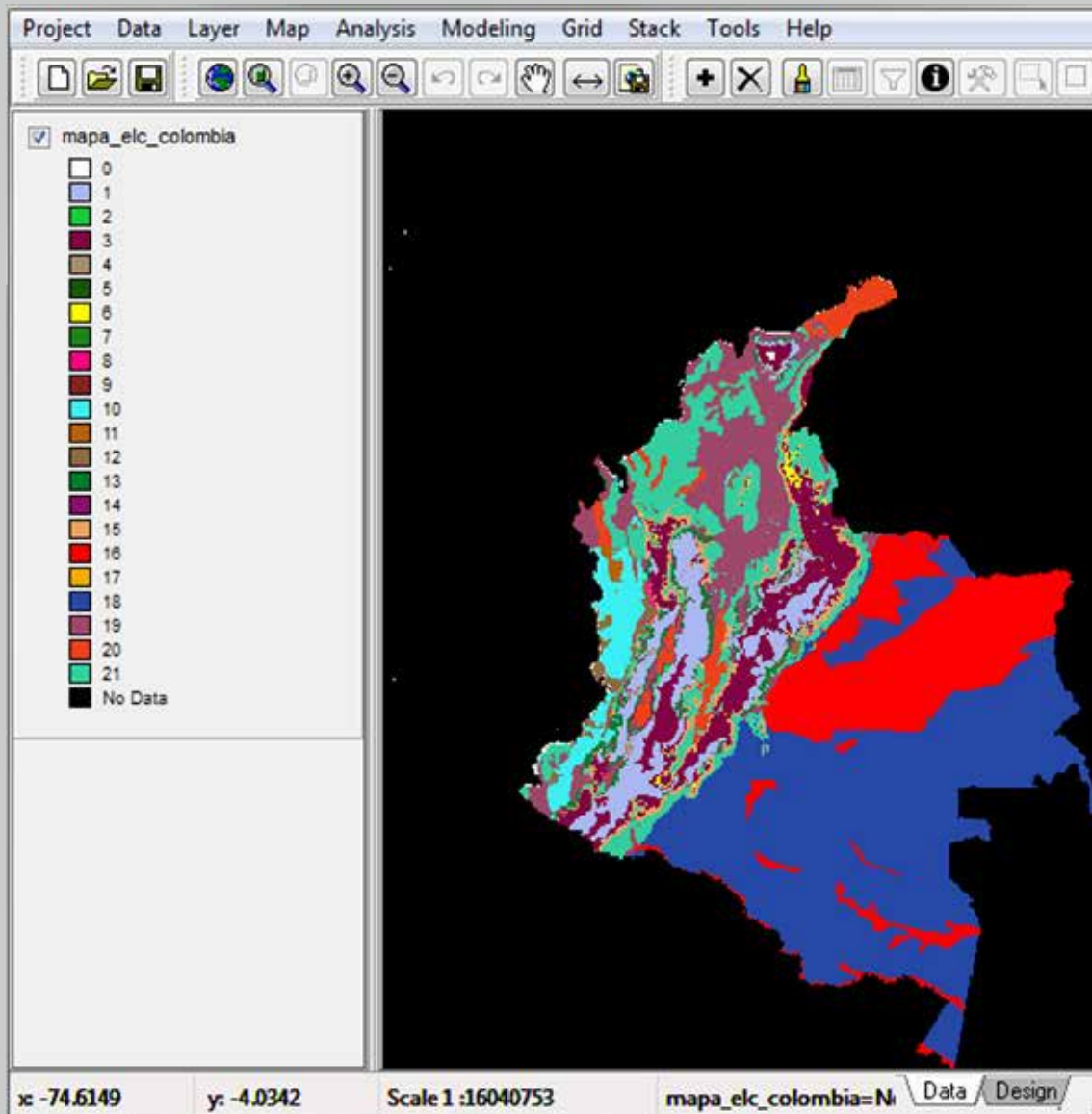
**FIGURE 17.** Example of a map of Colombia generated by the ELCmapas tool as pictured in DIVA-GIS before any editing. The 20 categories are shown grouped into 5 rows.



**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".



**FIGURE 18.** Example of a map of Colombia, generated by the ELCmapas tool, showing a color for each category. The properties of the map as opened by DIVA-GIS have been altered to show each category in a different color.



In addition to DIVA-GIS compatible maps, a Google Earth map is also generated: "mapa\_elc\_country.kml". If you have Google Earth installed in 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 colors) and it does not have optimal graphic quality.

## 4.6.2 The Tables

The tables generated by ELCmapas are in tab-delimited text format and can be opened in programs such as Excel, OpenOffice, or R. Normally, when you right-click on file name, the "Open with" option offers some of these programs if they are installed.

4.6.2.1 "Tabla\_ELC\_celdas\_country.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.

4.6.2.2 "numero\_categorias\_country.txt". This contains a simple count of the ecogeographic categories which have been generated and represented in the resulting ELC map (column "N\_ELC\_CAT") and the number of categories generated by each component.

4.6.2.3 "Estadist\_ELC\_country.txt", "Estadist\_BIOCLIM\_country.txt", "Estadist\_EDAPH\_country.txt" and "Estadist\_GEOPHYS\_country.txt". These tables provide the descriptive statistics (average, minimum value, maximum value and standard deviation) for each of the original variables involved in generating the ELC map and for the maps of each component (bioclimatic, geophysical and edaphic) represented in the ELC map. These tables are similar to the S2 supplementary table presented in order to describe the categories of the ELC map in Parra Quijano et al. (2012 A).

## 4.7. References

Kaufman, I. and Rousseeuw, P.J. 1987, Clustering by means of Medoids, in Statistical Data Analysis Based on the L1-Norm and Related Methods. Y. Dodge (eds), 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.; Draper, D.; Torres, E. and 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. and Turok, J. (ed.) *Crop wild relative conservation and use*. CAB International, Wallingford.

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. 2011 B. Evaluation and validation of ecogeographical core collections using phenotypic data. *Crop Science* 51:694-703

Parra-Quijano, M. Iriondo, J.M., Torres, M.E. 2012 A. 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. 2012 B. 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. 2012 C. Spatial and ecogeographic approaches for selecting genetic ecogeographic reserves in Europe. In: N. Maxted, M.E. Dulloo, B.V. Ford-Lloyd, L. Frese, J. Iriondo and MAA Pinheiro de Carvalho (ed.) *Agrobiodiversity Conservation: securing the diversity of crop wild relatives and landraces*. CABI, Wallingford, UK.

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.

Thormann, I. 2012. Applying FIGS to crop wild relatives and landraces in Europe. *Crop Wild Relative* 8 14:16. <http://www.pgrsecure.org/publications>

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





## 5. ECOGEO Tool

### 5.1. 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 analyze the abiotic component, classified according to three principal features which are often considered in studies of crop adaptation (Ceballos-Silva and Lopez-Blanco, 2003) and agricultural zoning (Williams et al., 2008):

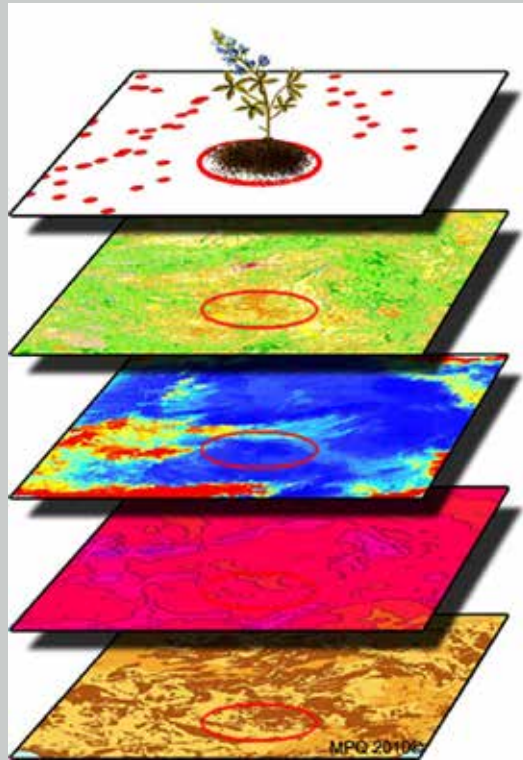
- (a) Bioclimatic: This refers to factors related to temperature and rainfall. It also includes the relationships between temperature and rainfall that are managed using indexes.
- (b) Geophysical: This brings together topographical and relevant relief factors, especially those related to solar radiation.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

Thus, 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, and, if considered in conjunction with other types of characterizations, such as phenotypic or genotypic, 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 cheap 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 are 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 the GEOQUAL tool should be used before performing a characterization of this type.

**FIGURE 19.** Process used to extract ecogeographic information for a collection site using GIS.



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 to determine 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 non-correlated variables describing the ecogeographic affinities between the inputs with a reduced number of components.

Please note that ecogeographic characterization yields information about the collection sites, rather than the nature of the germplasm itself. Therefore,

multivariate analyses which 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.

## 5.2. Characteristics of ECOGEO

The ECOGEO tool provides ecogeographic information for over 100 variables (see Annexes 12.1, 12.2 and 12.3) extracted for a list of accessions to be entered into the analysis using the format for passport data FAO/Bioversity 2012, with minor modifications. This format is used in several of the CAPFITOGEN tools.

The ECOGEO tool contains all the necessary information, meaning that the user does not have to download any information from the Internet. The information or layers of ecogeographic variables are adapted and arranged to work with the tool's R program settings.

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 the "country" parameter), the user can work with these areas of greater coverage, taking into account that the level of resolution of this information will probably be in a shorter range (cell sizes over 10x10 km).

Some considerations must be made concerning the way in which ecogeographic information is extracted from a collection site. Usually, extractions are performed at the point indicated by the coordinates. However, there are two situations when the "specific" extraction does not reflect the true nature of the abiotic conditions of the collection site:

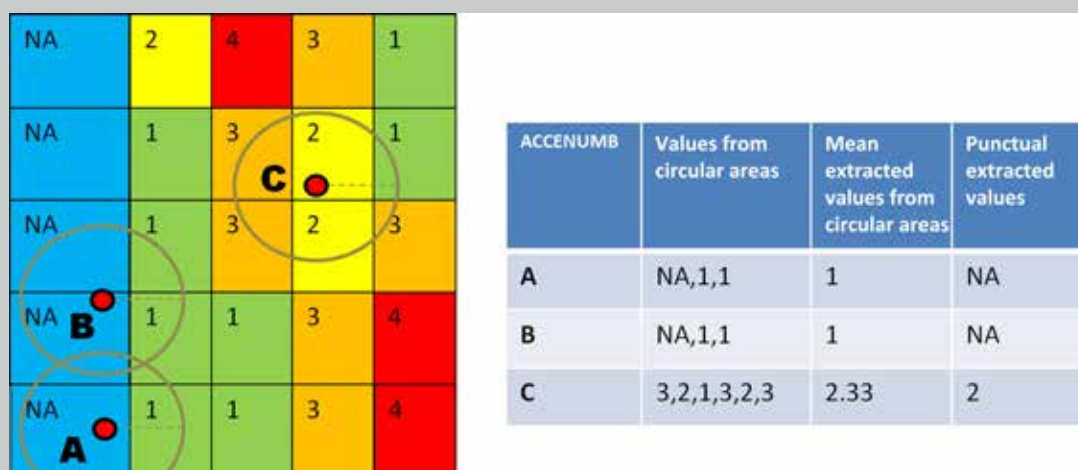
A) When there is little information available about the coordinates or these are poor quality, according to GEOQUAL or other methodologies. For instance, in cases of species with a coastal distribution, where, despite relatively accurate geo-referencing, a specific extraction may yield a number of "NA" values (no information available) because the ecogeographic information raster maps/layers do not mold themselves perfectly to 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 “radial” 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 (see Fig. 20). Additionally, ECOGEO automatically discards “NA” values when calculating and subsequently assigning values. To program ECOGEO to perform a “radial” extraction, the user must first activate the buffy parameter (see paragraph 5.3.1.6) and then enter the value in meters of the radius of the circular extraction area in the tamp parameter (paragraph 5.3.1.7).

**FIGURE 20.** Differences between the values assigned from a specific extraction point and a radial 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 on the basis of their coordinates.



Once the user has prepared the passport table according to the pre-established format, programmed the tool with the location, and indicated the work area, resolution and the extraction method required, the only remaining task is to define the variables/layers of interest for each aspect (bioclimatic, geophysical and edaphic) to characterize the germplasm collection sites.

With the definition of these parameters, in a single step, the 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 the “results” parameter.

Finally, if user is interested in performing a cluster analysis or a Principal Components Analysis (PCA), the tool can be programmed to run these analyses. The type of grouping and the number of main components to be retained may also be indicated at this point. The 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 the “results” parameter.

### 5.3. Using the ECOGEO Tool

Once the CAPFITOGEN tools have been installed and the GEOQUAL tool selected, you must define a set of parameters to ensure the R program runs correctly.

#### 5.3.1 Initial Parameters Defined by the User

##### 5.3.1.1 Parameter: *ruta*

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

##### 5.3.1.2 Parameter: *pais*

Explanation: Select the country where all or most of the data accessions you wish to analyze 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).

##### 5.3.1.3 Parameter: *pasaporte*

Explanation: 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 that this file must first be saved in the ‘Passport’ folder, which is part of the set of folders that make up the CAPFITOGEN directory.

##### 5.3.1.4 Parameter: *geoqual*

Explanation: Select this option if the passport data have been analyzed using the GEOQUAL tool and thus contain 50 columns (rather than the 45 columns in the passport model used by CAPFITOGEN tools). If so, please use the table generated

by GEOQUAL v.2 named "PasaporteOriginalEvaluadoGEOQUAL.txt" as a passport table in the point above.

*5.3.1.5 Parameter: totalqual*

Explanation: If your passport table is from GEOQUAL and you wish to set a minimum quality standard for your 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).

*5.3.1.6 Parameter: buffy*

Explanation: Check this option if you wish ecogeographic information to be extracted from an area around the collection site. Leaving this option unchecked means that information is extracted only from the point indicated by the collection site coordinates.

*5.3.1.7 Parameter: tamp*

Explanation: Specify the radius (in meters) 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 and cells without a value will not be taken into account. This value should not be lower than the distance from each side of the cell in the resol1 parameter.

*5.3.1.8 Parameter: resol1*

Explanation: Select the resolution level you wish to use to extract the ecogeographic information. Note that 1x1 km offers greater resolution but requires greater computing capacity and takes far longer than 5x5 km; however, this is not as limiting a factor as it is for the ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents or continents. See Annex 12.5 on the availability of resolutions in relation to the country or region selected.

*5.3.1.9 Parameter: bioclimsn*

Explanation: Select this option if you wish to include bioclimatic variables (temperature, rainfall and associated indexes).

*5.3.1.10 Parameter: bioclimv*

Explanation: Select the bioclimatic variables you wish to include in the ecogeographic characterization. All selectable variables are detailed in Annex 12.1.

#### 5.3.1.11 Parameter: *edaphsn*

Explanation: Select this option if you wish to characterize the information by soil variables (texture, depth, pH, etc.).

#### 5.3.1.12 Parameter: *edaphv*

Explanation: Select the edaphic variables you wish to include in the ecogeographic characterization. All selectable variables are detailed in Annex 12.2.

#### 5.3.1.13 Parameter: *geophyssn*

Explanation: Select this option if you wish to characterize the information by geophysical variables (related to terrain and sunlight).

#### 5.3.1.14 Parameter: *geophysv*

Explanation: Select the geophysical variables that you wish to include in the ecogeographic characterization. All selectable variables are detailed in Annex 12.3.

#### 5.3.1.15 Parameter: *latitud*

Explanation: Do you wish to include latitude as a variable of the ecogeographic characterization?

#### 5.3.1.16 Parameter: *longitud*

Explanation: Do you wish to include longitude as a variable of the ecogeographic characterization?

#### 5.3.1.17 Parameter: *ecogeoclus*

Explanation: Select this option if you wish to carry out an analysis of clusters of accessions by ecogeographic characterization.

#### 5.3.1.18 Parameter: *ecogeoclustype*

Explanation: Choose the type of hierarchical cluster to be used for ecogeographic clusters: "single" = nearest neighbor, "complete" = more compact neighborhood, "ward" = method of minimum variance of Ward, "mcquitty" = McQuitty's method, "average" = average similarity (UPGMA), "median" = similarity of the median, "centroid" = geometrically centroid, "flexible" = Beta flexible.

#### 5.3.1.19 Parameter: *ecogeopca*

Explanation: Select this option if you wish to perform an analysis of major components for accessions with an ecogeographic characterization.

#### 5.3.1.20 Parameter: *ecogeopcaxe*

Explanation: Number of components to be retained within the PCA analysis. This number should always be less than the number of ecogeographic variables.

#### 5.3.1.21 Parameter: *resultados*

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.

## 5.4. ECOGEO Results

In the path and folder created for “resultados” (parameter 5.3.1.21) two figures and four tables will be generated.

### 5.4.1 Figures

These are files called *dendrograma\_ecogeo.wmf* and *pca\_ecogeo.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 (parameter 5.3.1.17) or an analysis of the main components (parameter 5.3.1.19). They may be opened and even modified in Microsoft PowerPoint or image editing programs.

### 5.4.2 Tables

The four tables correspond to two types of information.

5.4.2.1 Ecogeographic characterization table of the accessions: This is the file called *TablaVarEcogeograficacountry.txt*. It corresponds to the initial characterization matrix and contains as many rows as accessions analyzed, and as many columns as ecogeographic descriptors.

5.4.2.2 Tables generated as a result of the analysis of main components: These correspond to the following files: *ecogeographic\_eigenvalues.txt* (table of eigenvalues), *ecogeographic\_eigenvectors.txt* (table of eigenvectors) and *ecogeographic\_pcascors.txt* (table containing each accession’s score for the main components retained). These are only generated if the tool has been required to perform this analysis (parameter 5.3.1.19).



---

## 5.5. References

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

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







## 6. Representa Tool

### 6.1. Concept of representativeness in germplasm collections

There are certain sensitive issues that 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. Basically, the idea is to calculate on a case-by-case basis the minimum number of individuals to be collected in order to ensure the capture of the majority 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:

$$GR = (NAC * 100) / NAT$$

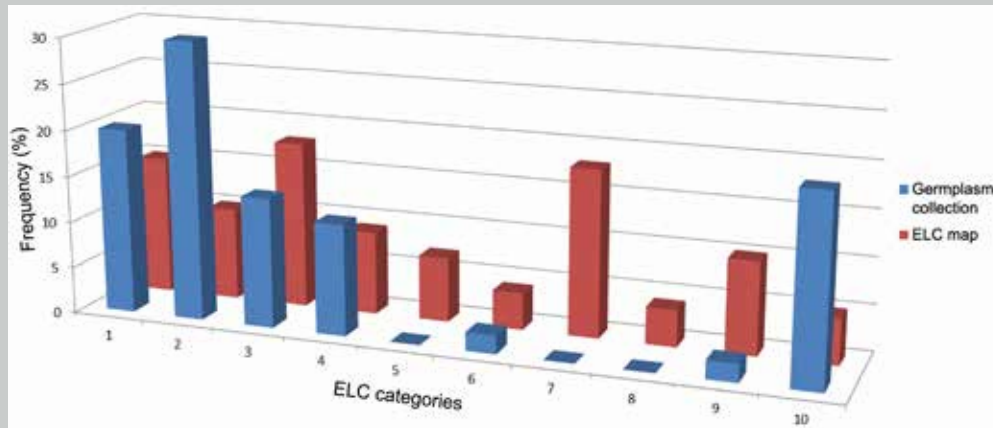
Where NAT is the total number of alleles in the sum of all the loci studied presented by the target species within the spatial area (continent, country, region, etc.) of collection to be evaluated, and NAC is the number of alleles of the loci of this species captured by this collection. This ideal determination of genetic representation entails a practical impediment. Knowing the total number of alleles which 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 which are known definitively to be composed of very few populations). Given the context of the plant genetic resources for food and agriculture, this exception is almost non-existent. 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 ELCmapas) to find out how many environmental conditions in a given spatial framework would be represented in a germplasm collection.

As an example of this application, Fig. 21 shows a frequency distribution for each ELC category of a germplasm collection contrasted with the availability of these categories in the total spatial framework. This fictitious example serves to show how the representativeness of a collection may, or may not, be biased according to the amount of environmental units present in the work area. The contrast between the values found in categories 2 and 7 reveals that the two distributions are highly dissimilar, and that it is possible that a Chi-squared test would determine an insignificant association between the two distributions.

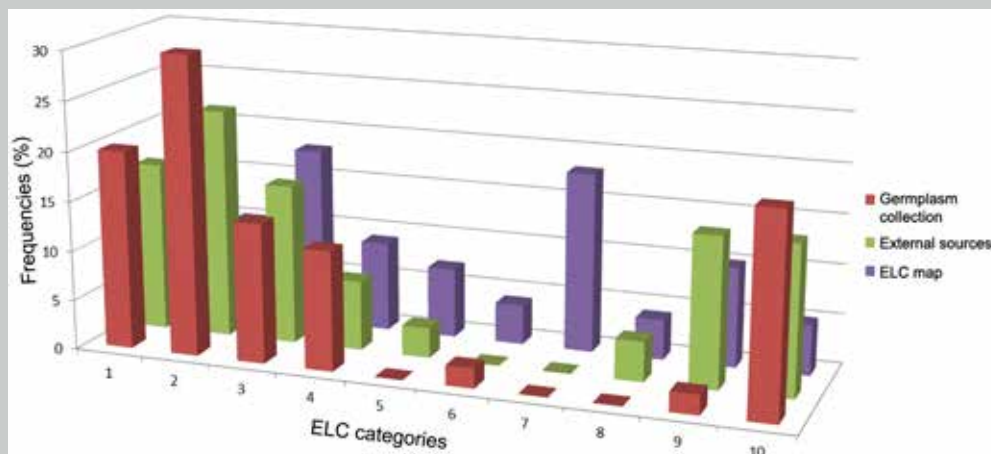
However, the most accurate determination of the ER is achieved using gap analysis. To do this, it is necessary to previously compile information from other sources external to the collection, such as other germplasm collections or any other data

**FIGURE 21.** 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).



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 which environments are under-represented in the collection.

**FIGURE 22.** 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. 22 illustrates the previously mentioned comparison process. Using the same fictional data from the example in Fig. 21, 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 because, thanks to external sources, the location of these populations is known.

It is important to make a clarification regarding how presence data from external sources could be analyzed. 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.

### 6.3. Using the Representa tool

Once the CAPFITOGEN tools have been installed and the Representa tool selected, it will be necessary to define a set of parameters to ensure the R program runs correctly.

#### 6.3.1 Initial Parameters Defined by the User

##### 6.3.1.1 Parameter: *ruta*

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

##### 6.3.1.2 Parameter: *internet*

Explanation: If you wish to download information about external sources from internet databases (GBIF) and have access to the Internet, please select this option.

##### 6.3.1.3 Parameter: *pasaporte*

Explanation: 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 "passport" folder which is part of the set of folders making up the CAPFITOGEN directory.

#### 6.3.1.4 Parameter: *geoqual*

Explanation: Select this option if the passport data have been analyzed using the GEOQUAL tool and thus contain 50 columns (rather than the 45 columns in the passport model used by CAPFITOGEN tools). Use the table from GEOQUAL v.2 called `PasaporteOriginalEvaluadoGEOQUAL.txt` as the passport table in the point above.

#### 6.3.1.5 Parameter: *totalqual*

Explanation: If your passport table is from GEOQUAL and you wish to set a minimum quality standard for your 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).

#### 6.3.1.6 Parameter: *fext*

Explanation: Do you have input from outside sources (meaning any information source other than the target collection being analyzed for representativeness) in the requisite format?

#### 6.3.1.7 Parameter: *fuentex*

Explanation: 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 text format and delimited by tabs). Please remember that this file should be saved in the Passport folder.

#### 6.3.1.8 Parameter: *geoqualfe*

Explanation: Does the table of externally-sourced input in the requisite format provide information about the quality of the geo-referencing in the required columns (prior application of GEOQUAL)?

#### 6.3.1.9 Parameter: *totalqualfe*

Explanation: If your table of externally-sourced input was evaluated by GEOQUAL and you wish to set a minimum quality standard to be met by the 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).

#### 6.3.1.10 Parameter: *duplibg*

Explanation: Check this option if you believe that input provided by outside sources from other databanks or germplasm collections is missing. (`TYPESOURCE` column with a value of 40). Please note that if you check this option, you may

make collections of populations that are already represented in other collections, leading to duplications between collections. Attention: Check this option if all occurrences of external sources that you are entering are from banks or germplasm collections. Otherwise, an error will be generated.

#### 6.3.1.11 Parameter: *gbifFE*

Explanation: Do you wish to download externally-sourced data from the Global Biodiversity Information Facility (GBIF) website? This option requires an Internet connection. This option is incompatible with the inclusion of externally-sourced input provided by the user. If you check this option and also provide a table with externally-sourced input, it will only take the latter into account.

#### 6.3.1.12 Parameter: *genero*

Explanation: Type the name of the genus of the species to be analyzed. This is the genus for which information will be downloaded from the GBIF website. Remember to capitalize the first letter. If the *gbifFE* parameter is not activated, it is not necessary to enter any information here.

#### 6.3.1.13 Parameter: *especie*

Explanation: Type the name of the species (epithet only) to be analyzed. This name will be placed next to the genus in order 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. If the *gbifFE* parameter is not activated, it is not necessary to enter any information here.

#### 6.3.1.14 Parameter: *mpaelc*

Explanation: Enter the name of the file containing the ELC map (generated by running the ELCmapas tool), which should be stored in the ELCmapas folder, one of the folders that makes up the CAPFITOGEN directory. 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".

#### 6.3.1.15 Parameter: *satelc*

Explanation: Enter the name of the file with the table of the ELC map's descriptive statistics generated using the ELCmapas tool (the tool usually names this file "Estadist\_ELC\_" plus the name of the country or region, corresponding to the result 4.6.2.3). Like the ELC map, this file should also be located in the ELCmapas folder. 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".

#### 6.3.1.16 Parameter: *dstdup*

Explanation: Determine the distance (in km) under which you consider two presence or collection sites to represent in fact the same population. The value zero (by default) excludes accessions with identical coordinates from the representativeness analysis.

#### 6.3.1.17 Parameter: *resultados*

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.

## 6.4. Results of Representa

In the path and folder created for “resultados” (parameter 6.3.1.17) up to five maps and up to five tables will be generated.

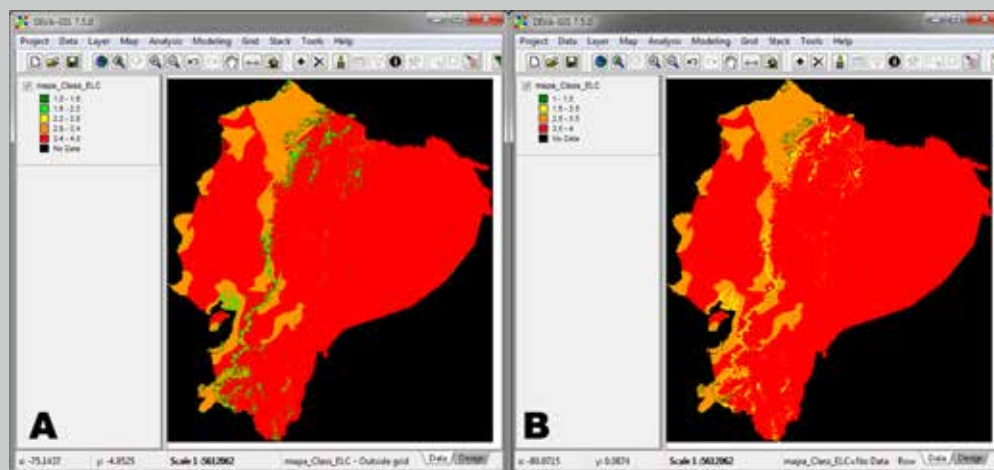
### 6.4.1 Maps

These are two vector point maps (shapefiles) and three raster maps (in the .grid format) that can be directly displayed in DIVA-GIS. If data from external sources are not introduced, there will be only three maps (as described in sections 6.4.1.1, 6.4.1.2 and 6.4.1.4).

6.4.1.1 “mapa\_Class\_ELC.grd”. This map divides the original ELC map categories into four groups (by quartile) according to their frequency across the whole territory. The frequency is divided on the basis of quartiles. Group 1 corresponds to the lowest frequency (below the 0.25 quartile); Group 2 is 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 is opened in DIVA-GIS, five colors are displayed by default, so it is important to change the display to ensure that only four colors are seen. Each color corresponds to a range representing each Group value (1 to 4). This is illustrated in Fig. 23.

6.4.1.2 “mapa\_Class\_Sp.grd”. This map divides the original ELC map categories into four groups (by quartile). These groups correspond to the division of categories by frequency across the whole territory. The frequency is divided on the basis of quartiles. Group 1 corresponds to the lowest frequency (below the 0.25 quartile); Group 2 is 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).

**FIGURE 23.** The appropriate visual configuration for Representa raster maps. A) Display of the *mapa\_Class\_ELC.grd* file as opened in DIVA-GIS. B) Display adjusted to four colors (one per frequency group). The least frequent adaptive scenarios (low and medium-low frequency groups) appear in green and yellow.



6.4.1.3 “*mapa\_Tipo\_faltante.grd*”. 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



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>
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 y NA	Not applicable	Not applicable

<sup>1</sup> This value is determined by comparing occurrences in external sources with those from germplasm collections/banks in each category according to the following formula:  $DIF = (FE/BG)/FE$ . FE refers to the number of occurrences from external sources while BG refers to the germplasm bank.

<sup>2</sup> This classification is the same as that shown in map 6.4.1.2.

<sup>3</sup> This classification is the same as that shown in map 6.4.1.1.

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 opens in DIVA-GIS it does not show the 13 classes with an individual color for each class, but all 13 values into five colors. The correct display is achieved using DIVA-GIS to add 8 more colors and adjusting the value ranges of each color (as in previous maps) to the value of a class.

6.4.1.4 "Shapefile\_Puntos\_BG.shp". Vector map (shafile) 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.

6.4.1.5 "Shapefile\_FE\_class.shp". Vector map (shafile) representing the occurrences from external sources. The table accompanying this points map presents the following fields in addition to the format data from external sources:

FE\_cat: Category of the ELC map where 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.

## 6.4.2 Tables

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

6.4.2.1 "Tabla\_Fuentes\_Externas\_clasificadas.txt". This corresponds to the same table accompanying the shapefile in paragraph 6.4.1.5, and contains the same variables.

6.4.2.2 "Tabla\_Resultados\_Representatividad.txt". This table presents the final results of the representativeness evaluation, whether or not data from external sources has been included. With this table, it is possible to create bar graphs in Excel as shown in Figs. 21 and 22. Finally, this table presents all the information required to calculate the parameters in Table 1, including the class value used to define priorities.

6.4.2.3 "Tabla\_Resultados\_X2.txt". 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 BG) vs. external sources (FE), and bank/collection vs. distribution of total frequencies of the ELC map categories.

6.4.2.4 "TablaClasificacionCuartilesEspecie.txt" and "TablaClasificacionCuartilesMapa ELC.txt". 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.

## 6.5. References

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

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

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

Parra-Quijano, M.; Draper, D.; Torres, E. and 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. and Turok, J. (ed.) Crop wild relative conservation and use. CAB International, Wallingford.

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



## 7. DIVmapas Tool

### 7.1. 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 estimate of parameters belonging to population genetics. However, in this case, before they are applied to all samples at once, diversity is estimated at local level with the determination of neighborhoods or areas of influence. The results of putting together all the results from each neighborhood 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 analyze 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 microsatellites (*Theobroma cocoa*), in addition to other analyses, in order to identify evolutionary processes in this cultivated plant.

On the basis of 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 clearly be replicated as the only element that varies is the genetic parameter which is calculated from the samples making up a neighborhood. Thus, if the parameter expresses the genetic differences between samples from a specific neighborhood, the map could be called a “diversity map”. The “DIVmapas” tool was developed on the basis of this methodology, and its application broadened beyond genotypic characterization data.

It is very important to note that these maps show genotypic diversity at the intra-specific level, one aspect that differentiates them notably from maps showing the wealth 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), are able to 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 concerning ex-situ and in situ conservation.

### 7.1.1 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”, because 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 on the basis of germplasm characterization data. Ecogeographic characterization is carried out by extracting information for each coordinate using a 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 and his team (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. Obviously, 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, 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 framework.

## 7.2. Procedure for Obtaining Diversity Maps Using the DIVmapas Tool

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

The 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

neighborhoods (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, which does not imply that it cannot be used in other biological fields. As a result, the 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 the 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. Section 3 of this manual refers to the GEOQUAL tool, which provides information on the quality of the geo-referencing of the germplasm collection site. It is advisable to use this tool before using the DIVmapas tool, so that only accessions with sufficiently high geo-referencing quality are taken into account when obtaining diversity maps. In any case, accessions without coordinates (DECLATITUDE and DECLONGITUDE or LATITUDE and LONGITUDE fields) will not be included in the analysis performed by the 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 “Formats” (Excel .xls files) folder. If you require an ecogeographic diversity map, please note that the DIVmapas tool includes the same germplasm ecogeographic classification process as the 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.

The DIVmapas tool will take advantage of all the valid characterization information available and accordingly create diversity maps for each individual 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. It is essential that identification codes for the accessions in the genotypic or phenotypic characterization tables be included in the FAO/Bioversity 2012 passport table containing geo-referencing information from the collection sites.

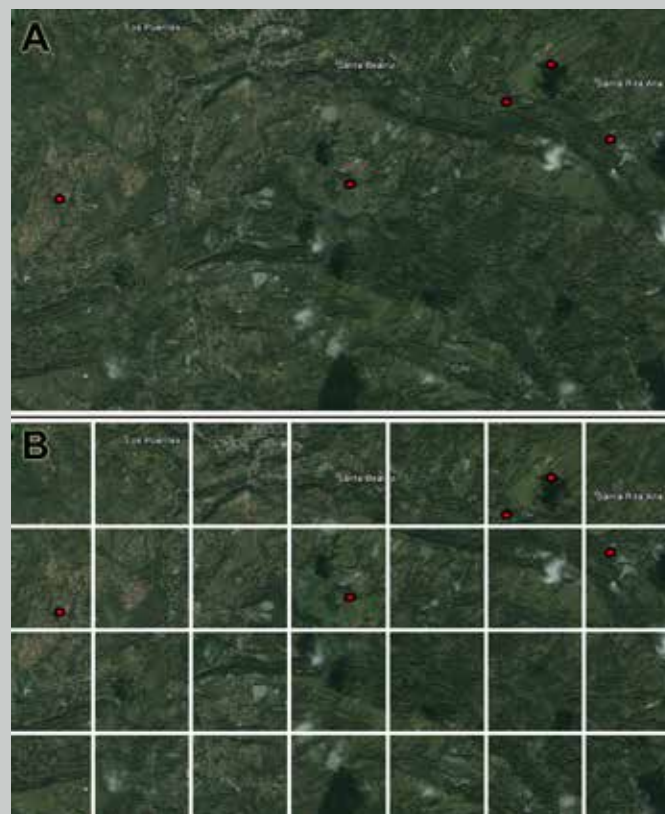
Once these conditions are clear, the following points show how the DIVmapas tool generates diversity maps, independently of the characterization data used for this purpose.

### 7.2.1. Distribution of collection sites and generation of grid

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 Figure 24). Additionally, a layer including the centroids of each cell in the grid is loaded (see Fig. 25 part A). Each centroid has an identification code.

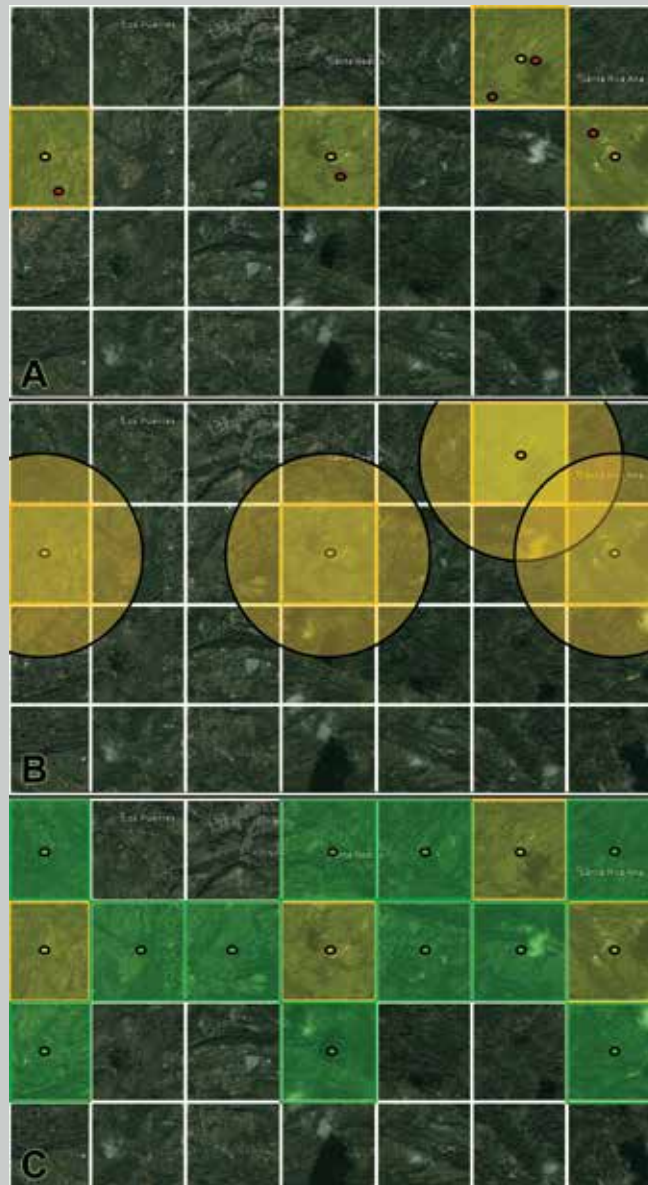
**FIGURE 24.** First step. A) spatial distribution of the collection sites, B) overlay of cell dimension grid (resolution) selected by the user.



### 7.2.2. Selection of cells with accessions and neighborhood 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 and 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 neighborhood cells, which are cells without accessions lying close to those initially selected (cells with accessions). For a cell to qualify as a neighborhood 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 neighborhood cells is shown in Fig. 25.

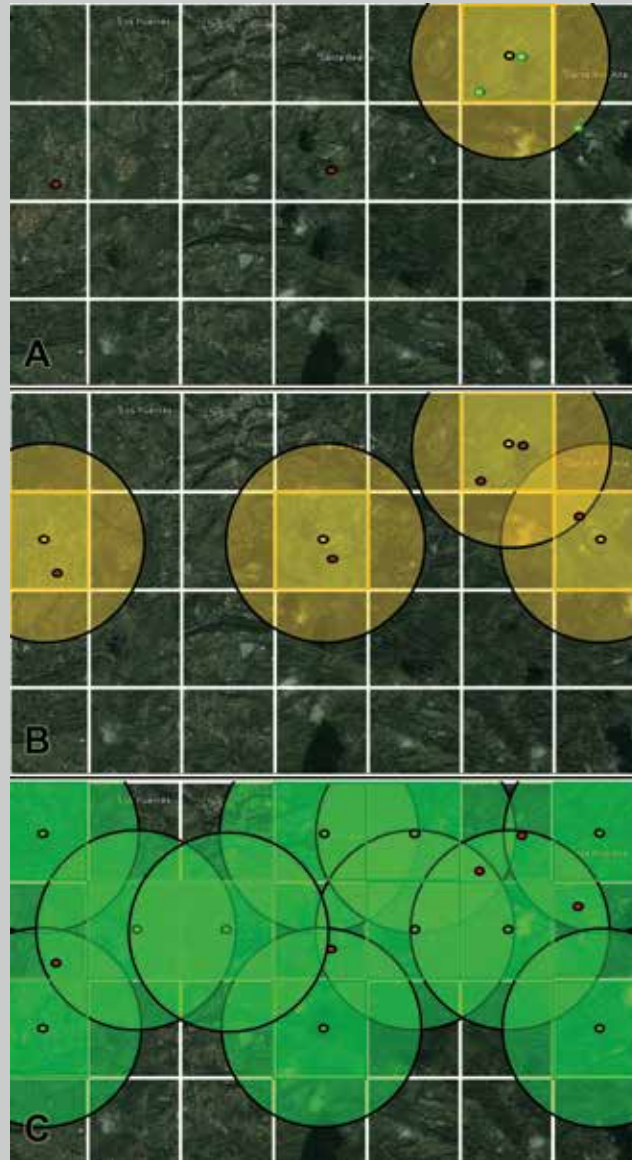
**FIGURE 25.** *Second step. A) Determination of cells with accessions and their centroids; B) projection of the areas of influence from the centroids of cells with accessions; C) determination of neighborhood cells.*



### 7.2.3. Determination of accessions linked to cells with accessions and neighborhood cells

The circular areas of influence are again projected from the centroids of the cells with accessions and the neighborhood cells. The ensuing list of accessions falling into each area is assigned an identification code for its respective centroid (see Fig. 26).

**FIGURE 26.** 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; C) determination of accessions occurring within the areas of influence of neighborhood cells.



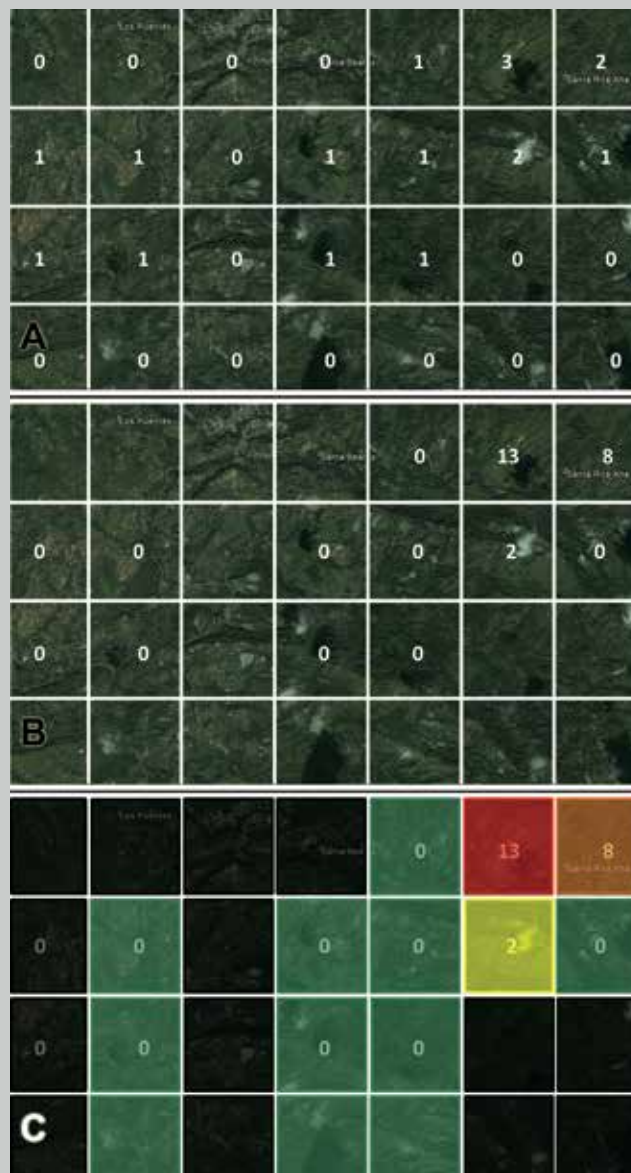
#### 7.2.4. Obtaining final diversity maps

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 neighborhood cells indicates the number



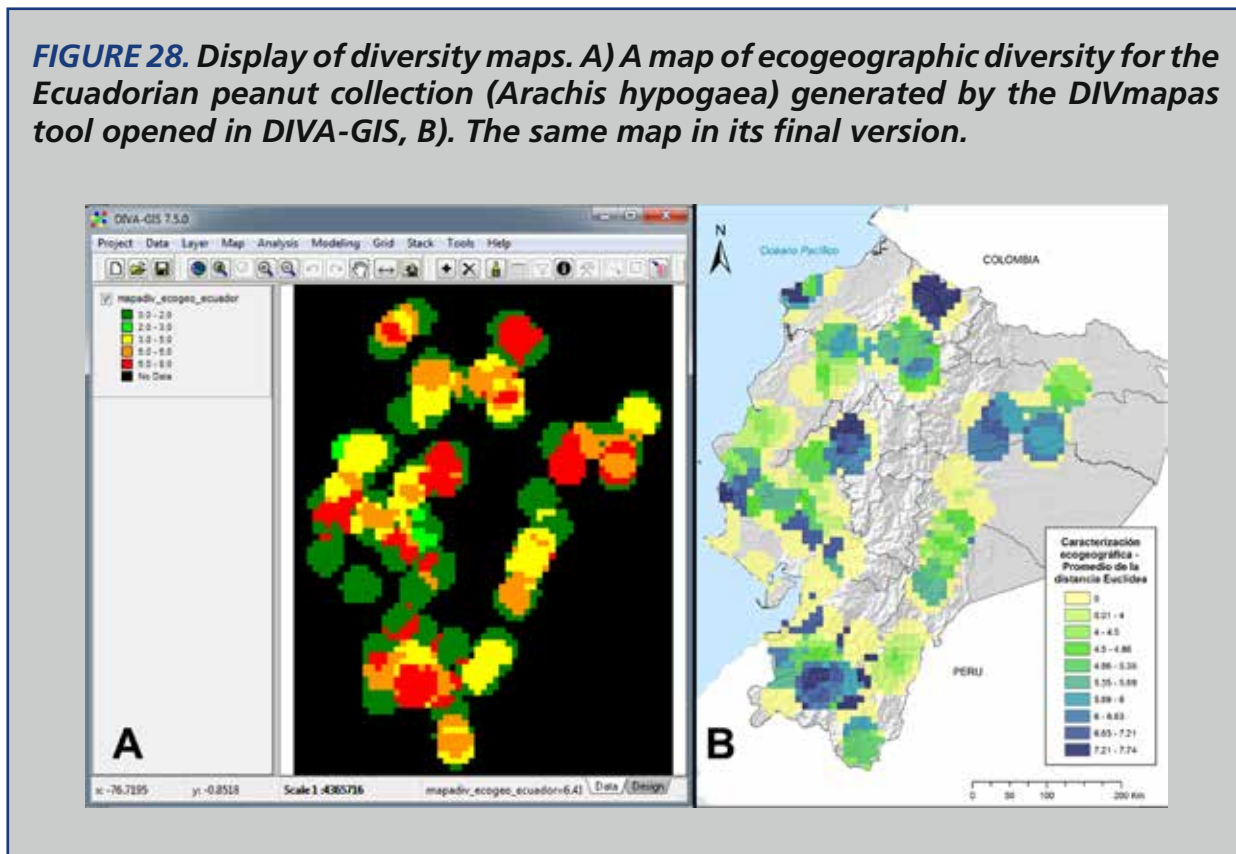
of initial matrices to be obtained. The process to standardize data is applied to each initial matrix when the data involves quantitative variables. Subsequently, a distance or similarity/dissimilarity coefficient is applied, which also produces a diagonal distance matrix. The average distance of the accessions included is calculated on the basis of this matrix and assigned to each centroid code and its respective cell. This allows R to produce raster cell maps reflecting the values assigned (see Fig. 27).

**FIGURE 27. Fourth step. A) The number of accessions analyzed by cell; B) values assigned to cells of an average genotypic, phenotypic or ecogeographic distance; C) assignment of colors graded according to the average values of distance.**



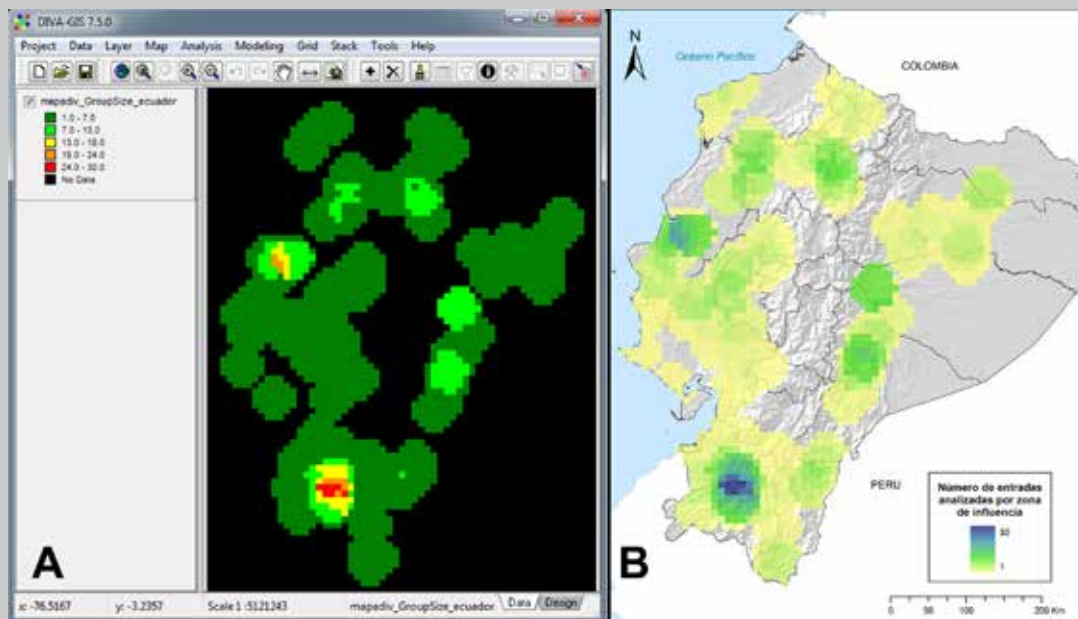
In the case of genotypic characterization, in addition to the average distance or dissimilarity, other genetic parameters may be calculated, such as Nei's measure of genetic diversity (1987), or the proportion of polymorphic markers for each group of accessions within each area of influence. R calculates these parameters using the initial characterization matrices.

**FIGURE 28.** Display of diversity maps. A) A map of ecogeographic diversity for the Ecuadorian peanut collection (*Arachis hypogaea*) generated by the DIVmapas tool opened in DIVA-GIS, B). The same map in its final version.



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 color from a graded color swatch. This allows you to quickly see the locations with the highest levels of diversity, as measured by the average values of distance/dissimilarity or by other genetic parameters (see Fig. 28). 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 analyzed by cell is also generated, as shown in Fig. 29, corresponding to Fig. 27 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 29.** Display of the map with the number of accessions analyzed by cell. A) Map opened in the DIVA-GIS program; B) the final version of the same map.



### 7.2.5. Use of resampling to eliminate any potential collection bias

Van Zonneveld and collaborators (2012) suggest using a method called rarefaction, which eliminates the effect of any potential bias in the collection of the samples analyzed. Thomas and collaborators (2013) suggest a resampling method. The latter method is incorporated into the DIVmapas tool.  $N$  (sample size) is defined as the median number of accessions analyzed per cell. Cells with a lower-than-average number of accessions are discarded. The value assigned to each cell thus corresponds to the average of the average distance values obtained in each resampling process. Depending on the number of resamplings selected by the user and the number of cells in the diversity map, the process may take minutes or hours.

### 7.2.6. Other analyses

The DIVmapas tool also permits other types of analysis to be performed, particularly when data characterization 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 the 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.

### 7.3. Formats for data entered into DIVmapas

To ensure that the DIVmapas tool works properly, enter the different kinds of information in the indicated formats. Usually, these formats are located in the "Formats" folder within the CAPFITOGEN tools' folders and files. Inside this folder you will find another with the name "Formatos DIVmapas" and within it five Excel files.

#### 7.3.1. Model of passport data

As in other CAPFITOGEN tools, the passport data must be entered using the FAO/Biodiversity 2012 format with minor modifications (file "ModeloDatosPasaporte\_FAO\_BIODIVERSITY\_2012.xls"). Since DIVmapas allows you to work with data evaluated on the basis of its geo-referencing quality using GEOQUAL, in addition to the normal passport data model, there is also a model with additional fields for the GEOQUAL evaluation values (file "ModeloDatosPasaporte\_FAO\_BIODIVERSITY\_2012\_conDatosGEOQUAL.xlsx"). However, the easiest way to use GEOQUAL-evaluated passport data is to directly load the table containing all the GEOQUAL 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 "Passport" folder in the CAPFITOGEN tools' folders and files.

#### 7.3.2. Phenotypic data model

When the contents of the phenotypic data format (file "ModeloDatosFenotipicos.xlsx") are displayed, a green column will appear (which must be filled in) called "ACCENUMB". 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 less 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 "PWEIGHT") 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, it is important that no name 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 the 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 "Passport" folder together with the other characterization data tables and the passport data table.

### 7.3.3. Model of table of the 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 7.3.2, 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 with filling 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 "Passport" folder in the same way as the other data accession tables.



### 7.3.4. Genotypic data model

As mentioned above, DIVmapas is a way of creating diversity maps on the basis of genotypic germplasm characterization which analyzes 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 which 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 must thus be encoded with values 0 and 1.

As with the phenotypic information, the DIVmapas only recognize 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 "Passport" folder together with the other characterization data tables and the passport data table.

## 7.4. Using the DIVmapas tool

Once the CAPFITOGEN tools have been installed and the GEOQUAL tool selected, it will be necessary to define a set of parameters to ensure that the R program runs correctly.

### 7.4.1 Initial Parameters Defined by User

#### 7.4.1.1 Parameter: *ruta*

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

#### 7.4.1.2 Parameter: *pais*

Explanation: Select the country where all or most of the data accessions you wish to analyze 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).

#### 7.4.1.3 Parameter: *bootstrap*

Explanation: Check this option if you wish to calculate the values for maps using the bootstrapping (resampling) technique. Please note that choosing this option will reduce the number of number of areas of analysis on the maps where the density of collection sites is lower.

#### 7.4.1.4 Parameter: *bootn*

Explanation: Specify the number of resamplings if you have chosen the bootstrapping technique option.

#### 7.4.1.5 Parameter: *replac*

Explanation: Mark this option if you wish to perform resampling with replacements.

#### 7.4.1.6 Parameter: *pasaporte*

Explanation: 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". Remember to save the file first in the "Passport" folder which is part of the set of folders making up the DIVmapas tool.

#### 7.4.1.7 Parameter: *geoqual*

Explanation: Select this option if the passport data have been analyzed using the GEOQUAL tool and thus contain 50 columns (rather than the 45 columns in the passport model used by CAPFITOGEN tools). If so, please use the table generated by GEOQUAL v.1.2 named "PasaporteOriginalEvaluadoGEOQUAL.txt" as the passport table in the point above.

#### 7.4.1.8 Parameter: *totalqual*

Explanation: If your passport table is from GEOQUAL and you wish to set a minimum quality standard for your 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).

#### 7.4.1.9 Parameter: *buffy*

Explanation: Check this option if you wish ecogeographic information to be extracted from an area around the collection site. Leaving this option unchecked means that information is extracted only from the point indicated by the collection site coordinates.

#### 7.4.1.10 Parameter: *tamp*

Explanation: Specify the radius (in meters) 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 and cells without a value will not be taken into account. This value should not be lower than the distance from each side of the cell in the *resol1* parameter.

#### 7.4.1.11 Parameter: *ecogeo*

Explanation: Select this option if you wish to obtain an ecogeographic diversity map.

#### 7.4.1.12 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 and takes far longer than 5x5 km; however, this is not as limiting a factor as it is for the ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents or continents. See Annex 12.5 on the availability of resolutions in relation to the country or region selected.

#### 7.4.1.13 Parameter: *bioclimsn*

Explanation: Select this option if you wish to use bioclimatic variables (temperature, rainfall and associated indexes) to create a map of ecogeographic diversity.

#### 7.4.1.14 Parameter: *bioclimv*

Explanation: Select the bioclimatic variables you wish to include to create a map of ecogeographic diversity. All selectable variables are detailed in Annex 12.1.

#### 7.4.1.15 Parameter: *edaphsn*

Explanation: Select this option if you wish to classify the information by soil variables (texture, depth, pH, etc.) to create maps of ecogeographic diversity.

#### 7.4.1.16 Parameter: *edaphv*

Explanation: Select the edaphic variables you wish to include to create a map of ecogeographic diversity. All selectable variables are detailed in Annex 12.2.

#### 7.4.1.17 Parameter: *geophyssn*

Explanation: Select this option if you wish to classify the information by geophysical variables (related to terrain and sunlight) to create maps of ecogeographic diversity.

#### 7.4.1.18 Parameter: *geophysv*

Explanation: Select the geophysical variables you wish to include to create a map of ecogeographic diversity. All selectable variables are detailed in Annex 12.3.

#### 7.4.1.19 Parameter: *latitud*

Explanation: Do you wish to include latitude on your map of ecogeographic diversity? Note: the inclusion of latitude and longitude will be reflected in distances on the final map, albeit at local level (within each area of influence).

#### 7.4.1.20 Parameter: *longitud*

Explanation: Do you wish to include longitude on your map of ecogeographic diversity? Note: the inclusion of latitude and longitude will be reflected in distances on the final map, albeit at local level (within each area of influence).

#### 7.4.1.21 Parameter: *phenotip*

Explanation: Select this option if you wish to obtain a map of phenotypic diversity. This map requires data accessions on phenotypic characterization or evaluation (e.g., morphology, phenology, productivity, resistance, etc.) in the format specified. Please remember to include the name of the file extension. For example, if the table is called "phenotypes", you should enter "phenotypes.txt" in this space. Remember that this table must be in the Passport folder stored in the CAPFITOGEN tools folder.

#### 7.4.1.22 Parameter: *phenot*

Explanation: Enter the name of the text file that contains data from the phenotypic characterization in the specified format. Please remember to include the name of the file extension. For example, if the table is called "phenotypes", you should enter "phenotypes.txt" in this space.

#### 7.4.1.23 Parameter: *phenotv*

Explanation: Enter the name of the text file that contains the table describing the nature of each phenotypic variable in the specified format. Please remember to include the name of the file extension. For example, if the table is called "variablesfenotipo", you should enter "variablesfenotipo.txt" in this space. This table should describe all the variables included in the characterization data table (see above).

#### 7.4.1.24 Parameter: *genotip*

Explanation: Select this option if you wish to obtain a map of genotypic diversity. This map requires data accessions on genotypic characterization or evaluation (such as the presence or absence of markers like zero and one) in the format specified. Remember that this table must be in the Passport folder stored in the CAPFITOGEN tools folder.

#### 7.4.1.25 Parameter: *genot*

Explanation: Enter the name of the text file containing the genotypic characterization data in the format specified. Please remember to include the name of the file extension. For example, if the table is called 'genotipos', you should enter 'genotipos.txt' in this space.

#### 7.4.1.26 Parameter: *neigd*

Explanation: 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 analyzed by cell.

#### 7.4.1.27 Parameter: *csimilar*

Explanation: Enter the similarity coefficient that you wish to use in order to generate the map of average genotypic distance. 1 = Jaccard index (1901), 2 = SMC by Sokal & Michener (1958), 3 = Sokal & Sneath (1963) (S5 by Gower & Legendre), 4 = Rogers & Tanimoto (1960), 5 = Dice (1945), 6 = Hamann coefficient, 7 = Ochiai (1957), 8 = Sokal & Sneath (1963) (S13 by Gower & Legendre), 9 = Pearson Phi coefficient, 10 = S2 by Gower & Legendre. Distance (d) is obtained as  $d = \sqrt{1-s}$  where s is the similarity coefficient.

#### 7.4.1.28 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).

#### 7.4.1.29 Parameter: *buffer*

Explanation: 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 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.

#### 7.4.1.30 Parameter: *ecogeoclus*

Explanation: Select this option if you wish to perform a cluster analysis for all accessions that include ecogeographical information.

#### 7.4.1.31 Parameter: *ecogeoclustype*

Explanation: Choose the type of hierarchical cluster to be used for ecogeographic clusters: "single" = nearest neighbor, "complete" = more compact neighborhood, "ward" = Ward's method of minimum variance, "mcquitty" = McQuitty's method,



"average" = average similarity (UPGMA), "median" = similarity of the median, "centroid" = geometrically centroid, "flexible" = Beta flexible.

#### 7.4.1.32 Parameter: *ecogeopca*

Explanation: Select this option if you wish to perform an analysis of major components for all accessions for which ecogeographic information has been extracted.

#### 7.4.1.33 Parameter: *ecogeopcaxe*

Explanation: Number of components to retain within the PCA analysis. This number should always be lower than the number of ecogeographic variables.

#### 7.4.1.34 Parameter: *phenoclus*

Explanation: Select this option if you wish to perform an analysis of clusters for all accessions including phenotypic information.

#### 7.4.1.35 Parameter: *phenoclustype*

Explanation: Choose the type of hierarchical cluster to be used for phenotypic clusters: "single" \ = nearest neighbor, "complete" \ = more compact neighborhood, "ward" \ = Ward's method of minimum variance, "mcquitty" \ = McQuitty's method, "average" \ = average similarity (UPGMA), "median" \ = similarity of the median, "centroid" \ = geometrically centroid, "flexible" \ = Beta flexible.

#### 7.4.1.36 Parameter: *phenopca*

Explanation: Select this option if you wish to perform an analysis of the main components/coordinates for all accessions including phenotypic information.

#### 7.4.1.37 Parameter: *phenopcaxe*

Explanation: Number of components/coordinates to retain within the PCA/PCoA analysis. This number should always be lower than the number of ecogeographic variables.

#### 7.4.1.38 Parameter: *phenovarq*

Explanation: Select this option if all the phenotypic variables/descriptors correspond to quantitative variables.

#### 7.4.1.39 Parameter: *genoclus*

Explanation: Select this option if you wish to perform a cluster analysis for all accessions that include genotypic information.

#### 7.4.1.40 Parameter: *genoclustype*

Explanation: Choose the type of hierarchical cluster to be used for genotypic clusters: "single" \ = nearest neighbor, "complete" \ = more compact neighborhood, "ward" \ = Ward's method of minimum variance, "mcquitty" \ = McQuitty's method, "average" \ = average similarity (UPGMA), "median" \ = similarity of the median, "centroid" \ = geometrically centroid, "flexible" \ = Beta flexible.

#### 7.4.1.41 Parameter: *genopco*

Explanation: Select this option if you wish to perform an analysis of the main coordinates for all accessions that include genotypic information.

#### 7.4.1.42 Parameter: *genopcoaxe*

Explanation: Number of components to retain within the PCoA analysis. This number should always be lower than the number of ecogeographic variables.

#### 7.4.1.43 Parameter: *mantelt*

Explanation: Please specify if you wish to analyze the correlation matrix (Mantel, 1967) between possible combinations of factors (ecogeographic vs. phenotypic vs. genotypic). All comparisons possible will be made according to whether phenotypic or genotypic data were entered or if an ecogeographic matrix was created on the basis of collection sites. A matrix of geographic distances will be generated for paired matrix comparisons.

#### 7.4.1.44 Parameter: *mantelmeth*

Explanation: Select the type of correlation to use for the Mantel test.

#### 7.4.1.45 Parameter: *mantelper*

Explanation: Enter as many permutations as desired for the Mantel test.

#### 7.4.1.46 Parameter: *resultados*

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.

## 7.5. Results of DIVmapas

When using DIVmapas, the number of graphs, tables and maps may vary according to the data entered, the options chosen and the analysis that the user wishes to perform. 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 the results parameter (see paragraph 7.4.1.46). The results are saved in the corresponding thematic folders explained in the following sections. The points map corresponding to the collection sites will be saved elsewhere in two versions ("ShapefilePuntosPasaporte.shp" and "mapa\_puntospas\_google.kml"). There is also a table ("Valor\_mediana\_para\_bootstrap.txt") which appears only when resampling processes are requested, and contains the average value used as a threshold for the process.

### 7.5.1 "ClassicMultivariateResults\_country" folder

This folder contains graphics (.wmf format) and tables (.txt) generated by multivariate analyses (cluster analysis and principal component analysis), as outlined in paragraphs 5.4.1 for figures and 5.4.2.2 for tables. 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). The results saved in this folder are generated by a simultaneous analysis of all accessions, i.e., the normal analytical procedure as performed by the ECOGEO tool.

### 7.5.2 "EcogeographicResults\_country" folder

This folder will appear when an ecogeographic diversity map is requested (see paragraph 7.4.1.11). It contains the diversity map in three different formats (as an image "mapadiv\_ecogeo\_country.png"; as a Google Earth map "mapadiv\_ecogeo\_country.kml"; and DIVA-GIS "mapadiv\_ecogeo\_country.grd"). In these maps, the average ecogeographic distances from each cell's accession of the area of influence are shown in different colors. This is the Euclidean distance, which has possible values ranging from 0 (when there is only one accession or all accessions were collected in identical environments) to infinity.

You will find the following tables:

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

7.5.2.2 "TablaVarEcogeograficascountry.txt". This table contains ecogeographic characterization data from the accessions analyzed. This table is similar to the one generated by the ECOGEO tool in paragraph 5.4.2.1.

7.5.2.3 "DistanciasMedias\_sin\_con\_bootstrap\_ECOGEO.txt". This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3). It

shows the average distance values for each cell (here denominated "cluster") with bootstrap ("W\_bootstrap") and without bootstrap (column "WO\_bootstrap").

7.5.2.4 "TestBootstrapping.txt". This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3). This table shows two association tests carried out between the average distance values obtained from processes performed with and without bootstrap (table 7.5.2.3).

### 7.5.3 "PhenotypicResults\_country" folder

This folder appears when a phenotypic diversity map is requested (see paragraph 7.4.1.11) and the table with the corresponding data has been entered. This contains the diversity map in three different formats (as an image "mapadiv\_phenot\_country.png"; as a Google Earth map "mapadiv\_phenot\_country.kml"; and DIVA-GIS "mapadiv\_phenot\_country.grd"). In these maps, the average phenotypic distances from each cell's accession of the area of influence are shown in different colors. 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 submitted have the same phenotype) up to 1 (maximum difference).

You will find the following tables:

7.5.3.1 "tabla\_estadisticas\_mapadiv\_phenot.txt". This table shows the statistics for the phenotypic diversity map in terms of distance, i.e., the average standard deviation and the maximum and minimum distance values defined in the set of cells making up the map.

7.5.3.2 "TestBootstrapping.txt". This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3). This table shows two association tests carried out between the average distance values obtained from processes performed with and without bootstrap.

### 7.5.4 "GenotypicResults\_country" folder

This folder appears when a genotypic diversity map is requested (see paragraph 7.4.1.11) and the table with the corresponding data has been entered.

Inside the folder you will find the following maps:

7.5.4.1 "mapadiv\_GenotDistance\_country". This corresponds to the map of genotypic diversity measured in average distances in three formats (image ".png"; Google Earth ".kml"; and DIVA-GIS ".grd"). In these maps, the average genotypic distances from the areas of influence of each cell are shown in different colors. The

distance corresponds to 1- Dice similarity coefficient (1945) and has potential values from 0 (when there is only one single accession or all accessions submitted have the same phenotype) to 1 (maximum difference).

7.5.4.2 "mapadiv\_GroupSize\_country". This corresponds to the map for the number of accessions analyzed by cell (".png" image; Google Earth ".kml", and DIVA-GIS ".grd"). In these maps, the number of accessions for the areas of influence of each cell are shown in different colors.

7.5.4.3 "mapadiv\_NeiGeneDiversity\_country". This corresponds to the map of genotypic diversity measured by Nei's diversity index (1987) in three formats (image ".png", Google Earth ".kml", and DIVA-GIS ".grd"). In these maps, the aforementioned diversity index obtained from the accessions characterized by the area of influence of each cell are shown in different colors.

7.5.4.4 "mapadiv\_ProportionVariableMarkers\_country". This corresponds to the map showing the proportion of polymorphic markers in three formats (image ".png", Google Earth ".kml", and DIVA-GIS ".grd"). 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 colors.

You will find the following tables:

7.5.4.5 "tabla\_estadisticas\_mapa\_GenotDistance.txt". This table shows the statistics for the ecogeographic diversity map (map 7.5.4.1) in terms of the Dice distance (1945), i.e., the average standard deviation and the maximum and minimum distance values defined in the set of cells making up the map.

7.5.4.6 "tabla\_estadisticas\_mapa\_NeiGeneDiversity.txt". This table shows the statistics for the genotypic diversity map (map 7.5.4.3) in terms of Nei's genetic diversity index (1987), i.e., the average standard deviation and the maximum and minimum distance values for this index, defined in the set of cells making up the map.

7.5.4.7 "NeiGeneDiversityMedias\_sin\_con\_bootstrap.txt". Table with Nei's genetic diversity indexes (1987) obtained for each cell (here called the "cluster") without bootstrapp ("WO\_bootstrap" column) and with bootstrap ("W\_bootstrap"). This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3).

7.5.4.8 "ProportPolymorphMarkersMedias\_sin\_con\_bootstrap.txt". Table showing the proportion of polymorphic markers obtained for each cell (here called the "cluster") without bootstrap ("WO\_bootstrap" column) and with



bootstrap (“W\_bootstrap”). This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3).

7.5.4.9 “DistanciasMedias\_sin\_con\_bootstrap\_DICE.txt”. This table shows Dice’s average distance values (1945) for each cell (here denominated “cluster”) without bootstrap (“W\_bootstrap”) and with bootstrap (column “WO\_bootstrap”). This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3).

7.5.4.10 “TestBootstrappingDICE.txt”. This table shows two association tests carried out between Dice’s average distance values (1945) obtained from processes performed with and without bootstrap. This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3).

7.5.4.11 “TestBootstrappingNei.txt”. This table shows two association tests carried out between Nei’s genetic diversity index (1987) values, obtained from processes performed with and without bootstrap. This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3).

7.5.4.12 “TestBootstrappingPPM.txt”. This table shows two association tests carried out between the proportion of polymorph markers obtained from processes performed with and without bootstrap. This table is only available when a bootstrap analysis has been requested (paragraph 7.4.1.3).

#### 7.5.4 “MantelCorrelationResults\_country” 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.

## 7.6. References

Damme, P., Garcia, W., Tapia, C., Romero, J., Manuel Siguéñas, M. and 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. 2012. FAO/Bioversity multi-crop Passport descriptors V.2. Available at:

[http://www.biodiversityinternational.org/index.php?id=19&user\\_biodiversitypublications\\_pi1%5BshowUid%5D=6901](http://www.biodiversityinternational.org/index.php?id=19&user_biodiversitypublications_pi1%5BshowUid%5D=6901)

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. and 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., and van Etten, J. 2012. Present spatial diversity patterns of *Theobroma cocoa* 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, Escibano 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



## 8. ColNucleo Tool

### 8.1. 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 which 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 which are as dissimilar as possible so that genetic duplicates or closely-related accessions are not included (Brown, 1995). Thus, in order 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 which many institutions are unable to afford. 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). This kind of collection does not however 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.

### 8.1.1 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 occur (using 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 and all that is required is to know to which cluster (ecogeographic map category) the new accession belongs (Parra Quijano et al., 2011b).

### 8.1.2 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 has 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):

- a) Random (R): Accessions are selected at random from the whole collection. Clusters created by stratification are ignored.
- b) Constant (C): The same number of accessions is selected from each cluster, regardless of how many accessions it contains.
- c) Proportional (P): The number of accessions selected from each cluster is proportional to its size (total number of accessions contained).
- d) Logarithmic (L): The number of accessions selected from each cluster is proportional to the logarithm of its size (total number of accessions contained).
- (e) 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.

### 8.1.3 Information about 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 which in practice cannot be created as the selected accessions are unavailable. There are several factors which influence an accession's availability for inclusion in a core collection, including the number of seeds available, or 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 for the curator to know what information is available when drawing up a core collection for practical purposes.

## 8.2. Ecogeographic Core Collections

Ecogeographic characterization is an alternative method of creating core collections. A core collection based on ecogeographic characterization, taking into account the relationship between phenotype, genotype and the environment, 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 Centre 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.

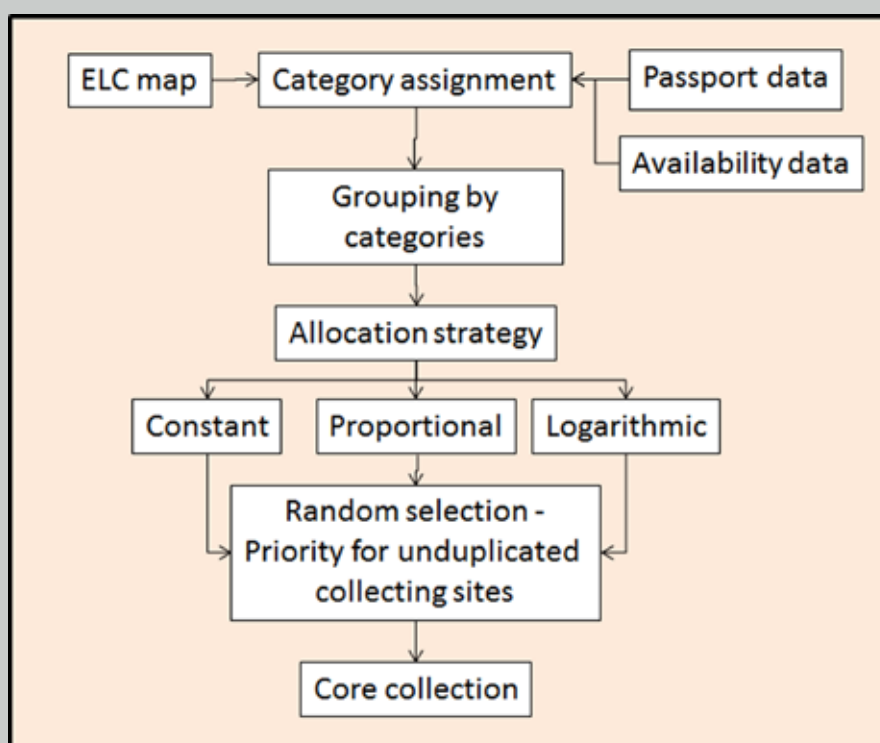
### 8.3. Obtaining ecogeographic core collections in ColNucleo

Following the recommendations of certain scientific studies on core collections and representativeness, the 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 4).

Fig. 30 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. If the quota is larger, all non-duplicate accessions will be selected and the shortfall 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 available 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 which accessions need to be made available in order for the core collection to represent all the ELC categories according to the quotas set.

**FIGURE 30.** Illustration of the process followed by the ColNucleo tool to obtain ecogeographic core collections.



#### 8.4. Format of passport table for ColNucleo

ColNucleo uses the FAO/Bioversity 2012 passport table with modifications which in turn uses the 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 the AVAILAB column, unavailable accessions with a 0 (zero) and those for which there is no information are coded with the letters NA.

#### 8.5. Using the ColNucleo Tool

Once the CAPFITOGEN tools have been installed and the ColNucleo tool selected, a set of parameters must be defined to ensure the R program runs correctly.

## 8.5.1 Initial Parameters Defined by User

### 8.5.1.1 Parameter: *ruta*

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

### 8.5.1.2 Parameter: *pasaporte*

Explanation: 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 "passport" folder which is part of the set of folders making up the CAPFITOGEN directory. 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 is an additional column which indicates the availability of each accession to be selected for a core/nuclear collection.

### 8.5.1.3 Parameter: *geoqual*

Explanation: Select this option if the passport data have been analyzed using the GEOQUAL tool and thus contain 51 columns (rather than the 46 in the passport model used exclusively by ColNucleo without having undergone GEOQUAL analysis). Therefore, please use the table generated by GEOQUAL called PasaporteOriginalEvaluadoGEOQUAL.txt as a passport table in the point above.

### 8.5.1.4 Parameter: *totalqual*

Explanation: If your passport table is from GEOQUAL and you wish to set a minimum quality standard for your 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).

### 8.5.1.5 Parameter: *mapaelc*

Explanation: Enter the name of the file containing the ELC map (generated by the ELCmapas tool), which should be found in the ELCmapas folder, one of the folders making up the CAPFITOGEN directory. This map should be in DIVA-GIS format (.grd extension, exactly as generated by the ELCmapas tool) and the name should be entered with the file extension. Thus, if the name of the map is "mapa\_elc\_spain", enter "mapa\_elc\_spain.grd".

### 8.5.1.6 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 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 ECLmapas folder. 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 called "Estadist\_ELC\_spain", it should be written "Estadist\_ELC\_spain.txt".

#### 8.5.1.7 Parameter: *distdup*

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites in fact represent the same population (geographical duplicate). The value of zero (which is the minimum and the default value) has a lower priority for the selection process for accessions with identical coordinates. The number of accessions considered to be geographical duplicates rises in tandem with increases in the distance value stipulated here.

#### 8.5.1.8 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".

#### 8.5.1.9 Parameter: *estratcol*

Explanation: Select a strategy in order 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 which are proportional to the number of accessions in each category); or "L" logarithmic (quotas which are proportional to the logarithm of the number of accessions in each category).

#### 8.5.1.10 Parameter: *availab*

Explanation: Select this option 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 1 (available), 0 (unavailable) or NA (no information/unavailable). If you prefer not to use information on availability, the selection of accessions will be carried out on the basis of the total number of accessions. Availability is defined according to the curator's own criteria and may be determined by the number of seeds preserved, their germination or a range of other factors.

#### 8.5.1.11 Parameter: *resultados*

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.



## 8.6. ColNucleo Results

Once the analysis is complete, ColNucleo produces three tables or, if the user has indicated that availability data should be used (parameter 8.5.1.10), four tables.

8.6.1.1 "CoreCollection.txt". This table contains the passport table with the accessions selected by ColNucleo for the ecogeographic core collection and an additional column on the far 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.

8.6.1.2 "CoreCollect\_Properties.txt". This table shows several parameters introduced by the user which ColNucleo has used to establish the core collection. The fields included are: "Allocation\_strategy"; "Sample\_size" (sampling intensity percentage); and "Use\_availability\_data" (whether availability data is used); "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).

8.6.1.3 "CoreCollect\_stats.txt". This table contains statistics for each ELC map category (identified in the "ELC\_CAT" column). It contains the following columns on the right of the ELC\_CAT column: "FREQ\_W\_DUPL" indicates the number of accessions, including geographical duplicates, whose collection site falls within each category; "FREQ\_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 as in paragraph 6.4.1.1.; "Duplicates" indicates the number of duplicate geographic accessions per category; "N\_Availab" indicates the number of total available accessions (duplicates and non-duplicates) per category; "N\_AvailabWO" indicates the number of non-duplicate accessions available per category; "Q\_Even" or "Q\_Prop" or "Q\_Log" (the column heading depends on the allocation method selected) refers to the quota (the number of accessions which each category should contain a priori); and lastly "CCfinal" indicates the number of accessions making up the ecogeographic core collection obtained by the ColNucleo tool on the basis of the parameters entered and (when applicable) availability of accessions.

8.6.1.4 "AccessionsToBeMultiplied.txt". This table has the same column structure as "CoreCollection.txt" except that it shows the accessions selected by ColNucleo as part of the core collection which are unavailable. It is assumed that these accessions need to be multiplied in order to become available for the core collection. However, there may be several reasons why they are unavailable, as explained in paragraph 8.1.3.

## 8.7. 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. (ed.) 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. *Aust. J. Agric. Res.* 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 pigeon pea (*Cajanus cajan* (L.) Millspaugh) using geographic and qualitative morphological descriptors. *Genetic Resources and Crop Evolution* 52:1049–1056.
- Tohme, J., P. Jones, S. Beebe, and M. Iwanaga. 1995. The combined use of agroecological and characterization 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. (ed.) 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. *Genet. Resour. Crop Evol.* 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. *Genet. Resour. Crop Evol.* 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. (ed.) Core collections of plant genetic resources. John Wiley & Sons, New York, NY.







---

## 9. FIGS\_R Tool

### 9.1. 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 ecogeographic information about collection sites and associations 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 its 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 its 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 plague *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 also 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 variable. 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 by 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 technique has a drawback in that it has to rely 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 in fact possess the trait of interest for which they were chosen using the ecogeographical conditions of their collection sites.

## 9.2. FIGS subsets and core collections

A FIGS subset is the set of accessions with potential for use in breeding a cultivated species and which 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

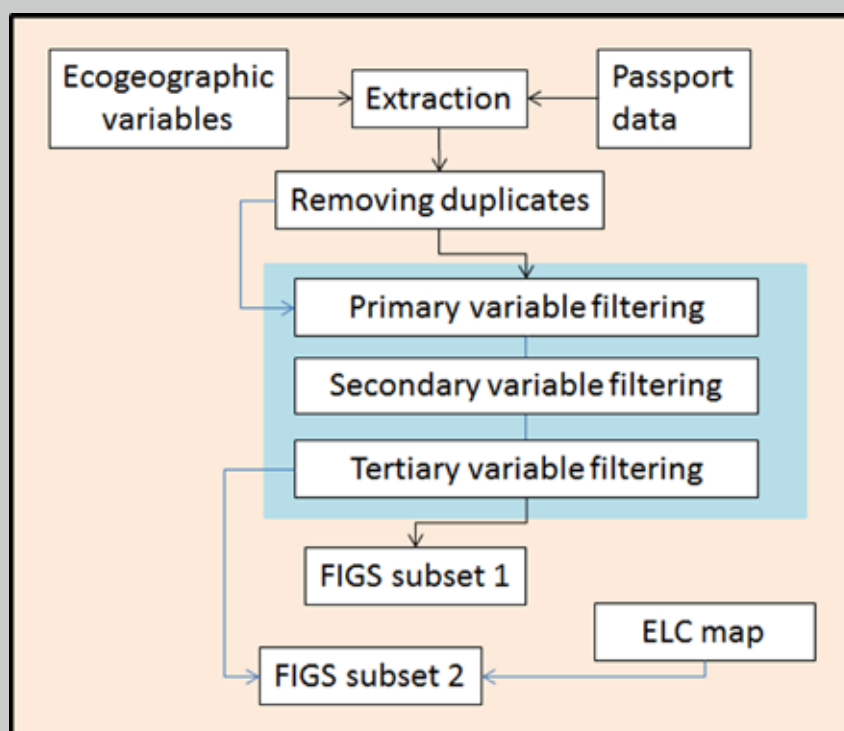


core collection is used to prioritize the characterization and evaluation of specific accessions in a collection when there are no resources to do this for the entire collection, a FIGS subset seeks to enhance the use of a germplasm collection, by helping crop breeders to locate material with the potential for integration into breeding programs.

### 9.3. Obtaining FIGS subsets using the FIGS\_R tool

The 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) which filters the subset generated by the second filter. Any one of the 103 ecogeographical variables can be chosen as the primary, secondary or tertiary variable (see Annexes 12.1, 12.2 and 12.3) available in the CAPFITOGEN tools. Fig. 31 shows the process followed by FIGS\_R to create FIGS subsets.

**FIGURE 31.** Illustration of the process followed by FIGS\_R to create FIGS subsets.



When using FIGS\_R, you can set the criteria for each selection variable. The first determines the range of values which the accession must meet in order to be included in the FIGS subset. The second determines a specific percentage 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).

In addition, the FIGS\_R tool can be used to create FIGS subsets which are ecogeographically balanced. In other words, if an ELC map has been created (with the ELCmapas tool, Chapter 4) 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 a greater ecogeographic representativeness which are still useful for breeding programs, given their trait of interest.

Finally, please note that the 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 "AVAILAB" field. To fill in the "AVAILAB" field, simply apply the criteria described in paragraph 8.4.

## 9.4. Using the FIGS\_R Tool

Once the CAPFITOGEN tools have been installed and the FIGS\_R tool selected, define a set of parameters to ensure the R program runs correctly.

### 9.4.1 Initial Parameters Defined by the User

#### 9.4.1.1 Parameter: ruta

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

#### 9.4.1.2 Parameter: *pais*

Explanation: Select the country where all or most of the data accessions you wish to analyze 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).

#### 9.4.1.3 Parameter: *pasaporte*

Explanation: 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. 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.

#### 9.4.1.4 Parameter: *geoqual*

Explanation: Select this option if the passport data have been analyzed using the GEOQUAL tool and thus contain 51 columns (rather than the 46 in the passport model used exclusively by ColNucleo without having undergone GEOQUAL analysis). Therefore, please use the table generated by GEOQUAL called PasaporteOriginalEvaluadoGEOQUAL.txt as a passport table in the point above.

#### 9.4.1.5 Parameter: *totalqual*

Explanation: If your passport table is from GEOQUAL and you wish to set a minimum quality standard for your 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).

#### 9.4.1.6 Parameter: *controlelc*

Explanation: Please specify 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 the user to obtain an additional FIGS subset in which accessions are selected for each ELC category. This requires the use of methods which make a selection from distribution fractions for all variables considered.

#### 9.4.1.7 Parameter: *mapaelc*

Explanation: Enter the name of the file containing the ELC map (generated by the ELCmapas tool), which should be found in the ELCmapas folder, one of the folders making up the CAPFITOGEN directory. This map should be in DIVA-GIS format (.grd extension, exactly as generated by the ELCmapas tool) and the name should be entered with the file extension. Thus, if the name of the map is "mapa\_elc\_spain", enter "mapa\_elc\_spain.grd".

#### 9.4.1.8 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 kind of file as "Estadist\_ELC\_" plus the name of the country or region). Like the ELC map, this file should be located in the "ELCmapas" folder. 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 called "Estadist\_ELC\_spain", enter "Estadist\_ELC\_spain.txt".

#### 9.4.1.9 Parameter: *distdup*

Explanation: Determine the distance (in km) under which you consider that two presence or collection sites in fact represent the same population (geographical duplicate). The value of zero (which is the minimum and the default value) has a lower priority for the selection process for accessions with identical coordinates. The number of accessions considered to be geographical duplicates rises in tandem with increases in the distance value stipulated here.

#### 9.4.1.10 Parameter: *availab*

Explanation: Select this option if you wish to use the column regarding the availability of accessions to be selected to make up the FIGS subset. This means prioritizing available accessions but not restricting the possibility of considering unavailable ones. 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 1 (available), 0 (unavailable) or NA (no information/unavailable). Availability is defined according to the curator's own criteria and may be determined by the number of seeds preserved, their germination or a range of other factors.

#### 9.4.1.11 Parameter: *soloavailab*

Explanation: Select this option 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 "AVAILAB" field).

#### 9.4.1.12 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 and takes far longer than 5x5 km; however, this is not as limiting a factor as it is for the ELCmapas tool. Resolutions of 10x10 and 20x20 may only be used for large countries, subcontinents or continents.

#### 9.4.1.13 Parameter: *buffy*

Explanation: Check this option if you wish ecogeographic information to be extracted from an area around the collection site. Leaving this option unchecked means that information is extracted only from the point indicated by the collection site coordinates.

#### 9.4.1.14 Parameter: *tamp*

Explanation: Specify the radius (in meters) 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.

#### 9.4.1.15 Parameter: *variab1v*

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

#### 9.4.1.16 Parameter: *variab1rang*

Explanation: Check this option if you wish to select accessions for the primary variable using a range of values, i.e., indicating minimum and maximum values of the range to use in selecting the accessions for the FIGS subset.

#### 9.4.1.17 Parameter: *variab1min*

Explanation: Specify the minimum value for the primary variable to determine the range to be used to select accessions for the FIGS subset.

#### 9.4.1.18 Parameter: *variab1max*

Explanation: Specify the maximum value for the primary variable to determine the range required to select accessions for the FIGS subset.



#### 9.4.1.19 Parameter: *variab1cola*

Explanation: Check this option 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.

#### 9.4.1.20 Parameter: *variab1vpor*

Explanation: 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.

#### 9.4.1.21 Parameter: *variab1vhl*

Explanation: Select the distribution fraction you wish to select for the primary variable.

#### 9.4.1.22 Parameter: *variab2*

Explanation: Check this option 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 which was previously selected using the primary variable.

#### 9.4.1.23 Parameter: *variab2v*

Explanation: Select one (1) secondary ecogeographic variable which you wish to use to select accessions for a FIGS subset. It may be the same as the primary variable.

#### 9.4.1.24 Parameter: *variab2rang*

Explanation: Check this option if you wish to select accessions for the secondary variable using a range of values, i.e., by indicating minimum and maximum values to set a range for selecting the accessions for the FIGS subset.

#### 9.4.1.25 Parameter: *variab2min*

Explanation: Specify the minimum value for the secondary variable in order to determine the range to be used to select accessions for the FIGS subset.

#### 9.4.1.26 Parameter: *variab2max*

Explanation: Specify the maximum value for the secondary variable in order to determine the range for selecting accessions for the FIGS subset.

#### 9.4.1.27 Parameter: *variab2cola*

Explanation: Check this option 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 with respect to the secondary variable.

#### 9.4.1.28 Parameter: *variab2vpor*

Explanation: 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.

#### 9.4.1.29 Parameter: *variab2vhl*

Explanation: Select the distribution fraction you wish to select for the secondary variable.

#### 9.4.1.30 Parameter: *variab3*

Explanation: Check this option 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 have no effect on the composition of a FIGS subset.

#### 9.4.1.31 Parameter: *variab3v*

Explanation: Select one (1) tertiary ecogeographic variable which you wish to use to select accessions for a FIGS subset. This may be the same as the primary or secondary variable.

#### 9.4.1.32 Parameter: *variab3rang*

Explanation: Check this option 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 which will be used to select accessions for a FIGS subset.

#### 9.4.1.33 Parameter: *variab3min*

Explanation: Specify the minimum value for the range of the tertiary variable to be used to select accessions for the FIGS subset.

#### 9.4.1.34 Parameter: *variab3max*

Explanation: Specify the maximum value for the range of the tertiary value to be used to select accessions for the FIGS subset.

#### 9.4.1.35 Parameter: *variab3cola*

Explanation: Check this option 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.

#### 9.4.1.36 Parameter: *variab3vpor*

Explanation: 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.

#### 9.4.1.37 Parameter: *variab3vhl*

Explanation: Select the distribution fraction you wish to select for the tertiary variable.

#### 9.4.1.38 Parameter: *resultados*

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

## 9.5. Results of FIGS\_R

Once the analysis is complete, FIGS\_R will produce between three and five tables according to the configuration of the parameters detailed above.

When an ELC map is not included in the analysis (parameter 9.4.1.6, *controlelc*), only the following three tables appear:

9.5.1 "FIGS\_regular.txt". This table identifies 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.

9.5.2 "FIGS\_stat\_table.txt". 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.

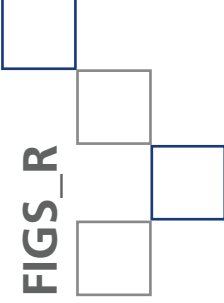
9.5.3 "Passport\_FIGS\_R.txt". This is the passport table introduced by the user into the analysis plus an additional field for each selection variable called "SEL\_VAR" followed by the numbers 1, 2 or 3. In these fields, accessions included in the FIGS subset are marked "1" while those not selected are marked "NA". Thus, the selection process using the primary variable (defined in parameter 9.4.1.15 ) selects the accessions identified with a "1" in the "SEL\_VAR1" field. If a secondary variable is used, the accessions selected during the second filtering process are identified with a "1" in the "SEL\_VAR2" field. Finally, if a tertiary variable is used, the accessions are identified with a "1" in the "SEL\_VAR3" field.

When an ELC map is included to provide more information about the ecogeographical characteristics of a FIGS subset, a new table appears:

9.5.4 "FIGS\_freq\_ELCmap.txt". This table shows frequency values as well as the number of duplicates and of accessions available for each ecogeographic category in a manner similar to the ColNucleo table described in paragraph 8.6.1.3 ("CoreCollect\_stats.txt"). 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 (in parameters 9.4.1.19, 9.4.1.27, and 9.4.1.35 respectively) – then the results will include a fifth table:

9.5.5 "FIGS\_UnderELC.txt". This is a table containing the same fields as in "Passport\_FIGS\_R.txt" (paragraph 9.5.3), but in this case it contains only those accessions from the FIGS collection 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 also have 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 ones and "var\_eco3" for the tertiary variable values).

The "FIGS\_freq\_ELCmap.txt" table (paragraph 9.5.4) 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.



## 9.6. References

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.







## 10. Frequent errors

The following list shows many of the error messages (text in red at the bottom of the interface), or at least the most frequent ones, which may occur when running tools. If other error messages appear when working with CAPFITOGEN tools or you have any questions about their operation, please visit the help forum at: <http://www.agrobiodiversidad.org/foroCAPFITOGEN/>.

### 1. No error message but no results in the folder designated for this purpose:

**Tool:** This can refer to any tool as the problem is due to the tool set being located in the wrong place.

**Solution(s):** Check that the data structure of the tools (the folders and files in the CAPFITOGEN folder) are all in a path with neither atypical values nor spaces. For example, if you saved the tools in the "Mis documentos" folder, the path could look like this: C:\Mis documentos\CAPFITOGEN. This path contains a space between the words "My" and "documents". This can generate an error where the analysis appears to have run successfully, but the folder where the results should be saved is empty. The CAPFITOGEN folder should be located in the root directory of the disk drive directly as follows: C:\CAPFITOGEN.

### 2. Error message:

**An error occurred: Error in the library(package name) : there is no package called 'package name'Calls: source -> withVisible -> eval -> eval -> library**

**Tool:** Any tool, as this is an installation issue.

**Solution(s):** The error indicates that one of the R packages requiring the tool 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 CAPFITOGEN) 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:

- a. 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 CAPFITOGEN 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 which refer to the version. Copy the file name and include the .zip extension.

- b. Open R using the executable hosted on the following path: C:\rwin\bin\i386\Rgui.exe, and type the following command in the “R console”:

```
Install.packages("X:/CAPFITOGEN/packages/nombreachivopaquete.zip")
```

X refers to the drive where the CAPFITOGEN tools are lodged (change this letter accordingly). Where it says “nombreachivo.zip”, paste the file name copied in step 1. Then hit “Enter”.

- c. The program will install the package and when it has finished, the following notification will appear:

```
package 'sp' successfully unpacked and MD5 sums checked
```

- d. Ensure that the package has been successfully installed, by typing:

```
library("package name")
```

“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”. A notification like the following will appear:

```
Lost warning notices
```

```
package 'cluster' was built under R version 2.15.3
```

- e. 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.

### 3. Error message:

**An error occurred: Error: unable to locate a vector of X.X Gb**

Tool (s): 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 ACCENUMB field. 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, assign each duplicate accession a unique number or code.

#### 4. Error message:

An error occurred: Error in sample.int(m, k) : first argument InvalidoCalls: source... withVisible -> eval -> eval -> kmeans -> sample.int

Tool(s): 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 "depth" variable often creates this problem. Using minimum rainfall variables in dry countries also tends to produce this problem

#### 5. 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

Tool(s): 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.

#### 6. 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

Tool(s): 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.



## 7. Error message:

An error occurred: Error: 'ecogeot' object not found

Tool(s): ECOGEO

Solution(s): Select the geophysv option if selecting geophysical variables.

## 8. 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

Tool(s): 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 (paragraph 6.3.1.10).

## 9. 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

Tool(s): 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 another error is generated regardless, use a lower resolution (larger cell size).

## 10. 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

Tool(s): ELCmapas

Solution(s): The error persists because the matrices generated are so large that the elbow method of determining the optimum number of clusters cannot manage them. The solution is to use a lower resolution (greater cell size).

## 11. Error message:

An error occurred: Error in .checkNumericCoerce2double(obj) : cannot retrieve coordinates from non-numeric elementsCalls: source ... coordinates -> .local -> do.call -> .checkNumericCoerce2double

Tool: 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.

## 12. Error message:

An error occurred: Error in readChar(con, 5L, useBytes = TRUE) : unable to open the connexionCalls: source -> withVisible -> eval -> eval -> load -> readChar

Tool: This may occur in any tool

Solution(s): This error usually corresponds to the wrong input of parameters. For example, in ELCmapas, this error may appear if you request that the tool use a cell resolution of 10x10 km for a country like Cuba. It also appears when the wrong path is entered for tools or passport tables, etc. To avoid this problem, check each parameter individually to ensure that the values are correct.

### 13. 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

Tool: 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 1,000 m using cell resolutions of 10x10 km approx. (5 arcmin). This will produce extraction values of zero and generate an error. Try using larger radii, ensuring they are greater than the size of side of each cell, and/or use a higher resolution. For example, if working with a radial extraction of 1,000 m, change "cells 5x5 km approx. (2.5 arcmin)" to "cells 1x1 km approx. (30 arcsec)" to solve the problem. If this does not work, try using specific extractions.

### 14. 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 <-

Tool: This may occur with tools where the user needs to enter passport details.

Solution(s): The error message may occur when, in the "passport" parameter, 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 which are not included. This can occur with ColNucleo, which expects the additional "AVAILAB" column. It can also occur if, under the "geoqual" parameter, the user indicates that the table has four extra columns containing the results of the GEOQUAL analysis and in fact it doesn't. Check the contents of the passport table you are entering and use the "geoqual" parameter accordingly.

### 15. 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 ó

An error occurred: Error in if (any(puntosorig\$DECLONGITUTE >= 180 | puntosorig\$DECLONGITUDE <= : value absent where TRUE/FALSE is necesarioCalls: source-> withVisible-> eval-> eval

Tool: This may occur with tools where the user needs to enter passport details.

Solution(s): There is an error in at least 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 and 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. These have extra rows which 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.







# 11. Acknowledgments

## 11.1 How to quote CAPFITOGEN

The following are the references required when quoting the use of CAPFITOGEN tools or the present user manual:

Parra-Quijano, M., Torres, E., Iriondo, J.M., López, F. 2014. CAPFITOGEN Tools User Manual Version 1.2 International Treaty on Plant Genetic Resources for Food and Agriculture, FAO, Rome.

## 11.2 Software used in CAPFITOGEN

The development of the CAPFITOGEN tools has been possible thanks to funding from the Ministry of Foreign Affairs and Cooperation of Spain (Ministerio de Asuntos Exteriores y de Cooperación de España) and the International Treaty on Plant Genetic Resources for Food and Agriculture.

CAPFITOGEN tools are supported by R version 2.15.2 (<http://cran.r-project.org/>).

R Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>

Rwui was used for the interface provided for GEOQUAL users (<http://sysbio.mrc-bsu.cam.ac.uk/Rwui/>).

Lextrend S.L developed the installer and improved the display interface (<http://www.lextrend.com/>).

## 11.3 R Packages

CAPFITOGEN tools use the following R packages:

SP (Edzer Pebesma, Roger Bivand, Barry Rowlingson, Virgilio Gomez-Rubio)  
raster (Robert J. Hijmans and Jacob van Etten)

maptools (Roger Bivand, Nicholas Lewin-Koh)

rgdal (Roger Bivand, Tim Keitt, Barry Rowlingson)

rgeos (Roger Bivand, Colin Rundel, Edzer Pebesma, Karl Ove Hufthammer)

RJSONIO (Duncan Temple Lang)

googleVis (Markus Gesmann, Diego de Castillo)

cluster (Martin Maechler)

modeltools (Torsten Hothorn, Friedrich Leisch, Achim Zeileis)

FPC (Christian Hennig)

dismo (Robert J. Hijmans, Steven Phillips, John Leathwick and Jane Elith)

ade4 (Daniel Chessel, Anne-Beatrice Dufour and Stephane Dray)

labdsv (David W. Roberts)

Vegan (Jari Oksanen, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B.) O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry, H. Stevens, Helene Wagner)

These packages in turn depend on other packages for their proper operation. We are grateful for the contributions of all the authors involved.

Some genotypic diversity maps using parameters derived from the AFLPdat scripts (<http://www.nhm.uio.no/english/research/ncb/aflpdat/>) of Dorothée Ehrich.

## 11.4 Data

The source for high precision administrative information is:

Global Administrative Areas GADM version 2 (<http://www.gadm.org>)

Maps with high-precision rings of 1 and 10 km around administrative areas (GADM) were obtained using the Buffer function (zone of influence) in ArcGIS 10.

Maps with low-precision rings of 1, 10 and 20 km are a modification of the world countries ESRI map (2011) under this license: Creative Commons Attribution-Noncommercial-Share Alike 3.0 United States License.

The Land Use Map used to calculate the SUITQUAL parameter is

Global Land Cover (GLC) 2000 and its reference is:

E. Bartholomé, A.S. Belward, F. Achard, S. Bartalev, C. Carmona-Moreno, H. Eva, S. Fritz;

J.M. Grégoire, P. Mayaux, H.J. Stibig (2002). Global Land Cover mapping for the year 2000

- Project status November 2002, Office for Official Publications of the European Communities,

Luxembourg EUR 20524).

The source of ecogeographic information (variables) is:

Worldclim (<http://www.worldclim.org>)

Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G. and Jarvis, A. 2005. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25:1965-1978.

Harmonized world soil database

(<http://webarchive.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/>)

FAO/IIASA/ISRIC/ISSCAS/JRC. 2012. Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Digital Elevation Models (DEM) of the Shuttle Radar Topography Mission (SRTM)

(<http://srtm.csi.cgiar.org/>)

Jarvis, A., H.I. Reuter, A. Nelson, E. Guevara, 2008, Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database: <http://srtm.csi.cgiar.org>.

## 11.5 Methodologies

The quality evaluation of georeferencing for passport data is a technique originally developed by the System for Ecogeographic Information for Spanish Plant Genetic Resources (Sistema de Información Ecogeográfica de los Recursos Fitogenéticos Españoles - SIERFE). The development of GEOQUAL, the Ecogeographic Land Characterization (ELC) Maps and the concept of Ecogeographic Representativeness (RE) are an original contribution by Mauricio Parra Quijano, Elena Torres Lamas and José María Iriondo Alegría.

The original concept used to develop diversity maps (DIVmapas tool) was published by:



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

The FIGS\_R tool incorporates ideas and developments achieved by PGR Secure (<http://www.pgrsecure.org>), a collaborative project funded under the Seventh Framework Programme (THEME KBBE 2010.1.1-03, "Characterization of biodiversity resources for wild crop relatives to improve crops by breeding). The concepts and developments introduced in FIGS\_R are from the work by the "Predictive characterization" group (Task 2.2) of the WP2 "Informatics", led by Bioversity International (<http://www.bioversityinternational.org/>). The researchers involved in these developments are: Imke Thormann, Jacob van Etten and Sonia Dias (Bioversity); José Iriondo and Luisa Rubio (Universidad Rey Juan Carlos); Shelagh Kell (University of Birmingham); Dag Endresen (GBIF); Rosa García (CRF-INIA); and Mauricio Parra Quijano (ITPGRFA).

## 11.6 Other Acknowledgments

Thanks to Fernando Latorre (CRF-INIA, Spain) for his firm support of the CAPFITOGEN program and the development of its tools. Thanks are also due for the support and comments supplied by other CRF-INRA researchers, in particular Lucía de la Rosa, Rosa García and Luis Ayerbe.

Thanks to Robert J. Hijmans for his valuable assistance in resolving issues with certain R packages and his generosity in permitting the distribution of worldclim and GADM information within the CAPFITOGEN tools.

We also thank many centers for research and conservation of plant genetic resources and their researchers and curators for their valuable comments and inputs to help improve the CAPFITOGEN tools.





## 12. Annexes

### 12.1 Available Ecogeographical Variables. Bioclimatic variables.

Code	Description of variables	Unit	Source
prec_1	Average rainfall for January	mm	Worldclim
prec_2	Average rainfall for February	mm	Worldclim
prec_3	Average rainfall for March	mm	Worldclim
prec_4	Average rainfall for April	mm	Worldclim
prec_5	Average rainfall for May	mm	Worldclim
prec_6	Average rainfall for June	mm	Worldclim
prec_7	Average rainfall for July	mm	Worldclim
prec_8	Average rainfall for August	mm	Worldclim
prec_9	Average rainfall for September	mm	Worldclim
prec_10	Average rainfall for October	mm	Worldclim
prec_11	Average rainfall for November	mm	Worldclim
prec_12	Average rainfall for December	mm	Worldclim
tmean_1	Average temperature for January	°C	Worldclim
tmean_2	Average temperature for February	°C	Worldclim
tmean_3	Average temperature for March	°C	Worldclim
tmean_4	Average temperature for April	°C	Worldclim
tmean_5	Average temperature for May	°C	Worldclim
tmean_6	Average temperature for June	°C	Worldclim
tmean_7	Average temperature for July	°C	Worldclim
tmean_8	Average temperature for August	°C	Worldclim
tmean_9	Average temperature for September	°C	Worldclim
tmean_10	Average temperature for October	°C	Worldclim
tmean_11	Average temperature for November	°C	Worldclim
tmean_12	Average temperature for December	°C	Worldclim
tmin_1	Minimum temperature for January	°C	Worldclim
tmin_2	Minimum temperature for February	°C	Worldclim
tmin_3	Minimum temperature for March	°C	Worldclim
tmin_4	Minimum temperature for April	°C	Worldclim
tmin_5	Minimum temperature for May	°C	Worldclim
tmin_6	Minimum temperature for June	°C	Worldclim
tmin_7	Minimum temperature for July	°C	Worldclim
tmin_8	Minimum temperature for August	°C	Worldclim

## 12.1 Continued

Code	Description of variables	Unit	Source
tmin_9	Minimum temperature for September	°C	Worldclim
tmin_10	Minimum temperature for October	°C	Worldclim
tmin_11	Minimum temperature for November	°C	Worldclim
tmin_12	Minimum temperature for December	°C	Worldclim
tmax_1	Maximum temperature for January	°C	Worldclim
tmax_2	Maximum temperature for February	°C	Worldclim
tmax_3	Maximum temperature for March	°C	Worldclim
tmax_4	Maximum temperature for April	°C	Worldclim
tmax_5	Maximum temperature for May	°C	Worldclim
tmax_6	Maximum temperature for June	°C	Worldclim
tmax_7	Maximum temperature for July	°C	Worldclim
tmax_8	Maximum temperature for August	°C	Worldclim
tmax_9	Maximum temperature for September	°C	Worldclim
tmax_10	Maximum temperature for October	°C	Worldclim
tmax_11	Maximum temperature for November	°C	Worldclim
tmax_12	Maximum temperature for December	°C	Worldclim
bio_1	Annual average temperature	°C	Worldclim
bio_2	Average daytime temperature range	°C	Worldclim
bio_3	Isothermality (BIOCLIM2/BIOCLIM7)(*100)		Worldclim
bio_4	Temperature seasonality (standard deviation*100)		Worldclim
bio_5	Maximum temperature for the warmest month	°C	Worldclim
bio_6	Minimum temperature for the coldest month	°C	Worldclim
bio_7	Annual temperature range (BIOCLIM5 - BIOCLIM6)	°C	Worldclim
bio_8	Average temperature for the quarter with most rainfall (3 rainiest months)	°C	Worldclim
bio_9	Average temperature for the driest quarter (3 driest months)	°C	Worldclim
bio_10	Average temperature for the hottest quarter (3 hottest months)	°C	Worldclim
bio_11	Average temperature for the coldest quarter (3 coldest months)	°C	Worldclim
bio_12	Annual rainfall	mm	Worldclim
bio_13	Rainfall during the wettest month	mm	Worldclim
bio_14	Rainfall during the driest month	mm	Worldclim
bio_15	Seasonality of rainfall (variation coefficient)	mm	Worldclim



## 12.1 Continued

Code	Description of variables	Unit	Source
bio_16	Rainfall during the wettest quarter (3 rainiest months)	mm	Worldclim
bio_17	Rainfall during the driest quarter (3 driest months)	mm	Worldclim
bio_18	Rainfall during the hottest quarter (3 hottest months)	mm	Worldclim
bio_19	Rainfall during the coldest quarter (3 coldest months)	mm	Worldclim

## 12.2 Available Ecogeographical Variables. Edaphic variables.

Code	Description of variables	Unit	Source
ref_depth	Depth reference for the soil unit	m	HWS Database
t_gravel	Gravel content in surface soil	%vol.	HWS Database
t_sand	Sand content in surface soil	% peso	HWS Database
t_silt	Silt content in surface soil	% peso	HWS Database
t_clay	Clay content in surface soil	% peso	HWS Database
t_ref_bulk	Apparent bulk density reference in surface soil	kg/dm <sup>3</sup>	HWS Database
t_oc	Organic carbon content in surface soil	% peso	HWS Database
t_ph_h2o	Surface soil pH in a soil-water solution	-log(H <sup>+</sup> )	HWS Database
t_cec_clay	Clay cation exchange capacity in surface soil	cmol/kg	HWS Database
t_cec_soil	Cation exchange capacity in surface soil (general)	cmol/kg	HWS Database
t_bs	Saturation of bases in surface soil.	%	HWS Database
t_teb	Total exchangeable bases in surface soil	cmol/kg	HWS Database
t_caco3	Calcium carbonate in surface soil	% peso	HWS Database
t_caso4	Gypsisols in surface soil	% peso	HWS Database
t_esp	Sodicity of surface soil	%	HWS Database
t_ece	Salinity of surface soil	dS/m	HWS Database
s_gravel	Gravel content of subsoil	%vol	HWS Database

## 12.2 Continued

Code	Description of variables	Unit	Source
s_sand	Sand content of subsoil	% peso	HWS Database
s_silt	Silt content of subsoil	% peso	HWS Database
s_clay	Clay content of subsoil	% peso	HWS Database
s_ref_bulk	Apparent bulk density reference in subsoil	kg/dm <sup>3</sup>	HWS Database
s_oc	Content of organic carbon in subsoil	% peso	HWS Database
s_ph_h2o	pH in subsoil in soil-water solution	-log(H <sup>+</sup> )	HWS Database
s_cec_clay	Clay cation exchange capacity in subsoil	cmol/kg	HWS Database
s_cec_soil	Cation exchange capacity in subsoil (general)	cmol/kg	HWS Database
s_bs	Saturation of bases in subsoil.	%	HWS Database
s_teb	Total exchangeable bases in subsoil	cmol/kg	HWS Database
s_caco3	Calcium carbonate in subsoil	% peso	HWS Database
s_caso4	Gypsisols in subsoil	% peso	HWS Database
s_esp	Sodicity in subsoil	%	HWS Database
s_ece	Salinity in subsoil	dS/m	HWS Database

## 12.3 Ecogeographic variables available. Geophysical variables.

Code	Description of variables	Unit	Source
alt	Altitude. Meters above sea level	m	Worldclim
slope	Gradient (in degrees) of the land surface	°	SRTM MDE
aspect	Orientation (in degrees) of the land surface	°	SRTM MDE
northness	Northness. 1 if it faces northwards, - 1 if it faces southwards		SRTM MDE
eastness	Eastness. 1 if it faces eastwards, - 1 if it faces westwards		SRTM MDE
POINT_X	Longitude	°	
POINT_Y	Latitude	°	

Note: The websites of information sources (worldclim, SRTM MDE and HWS Database) appear in Chapter 11 (Acknowledgments).



## 12.4 Explanation of the extra columns in the results table “tabla\_de\_analisisGEOQUAL.txt”.

Variable	Explanation
globlandc	Value extracted from GLC 2000 (Global Land Cover 2000).
DISTOLAND	Ring of distance within which the coordinates are found. (0 = ground, 1 = 1 km, 10 = 10 km, etc.).
SUITQUAL	SUITQUAL parameter (values from 0 to 20).
ID_0	Value extracted from GADM identifying the area of the country.
ISO	Value extracted from GADM compared with ORIGCTY.
NAME_0	Value extracted from GADM for the country's full name.
ID_1	Value extracted from GADM identifying the area at the NAME_1 level.
NAME_1	Value extracted from GADM compared with ADM1.
VARNAME_1	Value extracted from GADM for alternative names to NAME_1.
ENGTYPE_1	Value extracted from GADM defining the type of administration represented by NAME_1.
ID_2	Value extracted from GADM identifying the area at the NAME_2 level.
NAME_2	Value extracted from GADM compared with ADM2.
VARNAME_2	Value extracted from GADM for alternative names to NAME_2.
ENGTYPE_2	Value extracted from GADM defining the type of administration represented by NAME_2.
ID_3	Value extracted from GADM identifying the area at the NAME_3 level.
NAME_3	Value extracted from GADM compared with ADM3.
VARNAME_3	Value extracted from GADM for alternative names to NAME_3.
ENGTYPE_3	Value extracted from GADM defining the type of administration represented by NAME_3.
ID_4	Value extracted from GADM identifying the area at the NAME_4 level.
NAME_4	Value extracted from GADM compared with ADM4.
VARNAME_4	Value extracted from GADM for alternative names to NAME_4.

Variable	Explanation
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
errores	COORQUAL errors sub-parameter
precis	COORQUAL precis sub-parameter
georble	COORQUAL georble sub-parameter
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).

## 12.5 Table of cell size availability by region/country

To-date, ecogeographic information adapted to the CAPFITOGEN tools is available for 162 countries and 2 regions. The cell size available for countries is 30 arc-seconds (~1x1 km at the equator) and 2.5 arc-minutes (~5x5 km at the equator). In some large countries (such as Brazil), the ELCmapas tool may generate errors when using high-resolution information (1x1 km), although other tools are unlikely to encounter any problems.

As regards regions, there is ecogeographic information for Europe at 2.5 arc-minutes (~5x5 km at the equator) and for the world in two resolutions, 5 arc-minutes (~10x10 km) and 10 arc-minutes (~20x20 km). Problems with the ELCmapas tool may arise when using a 10x10 km resolution with countries larger than 1 million square kilometers.

Ecogeographic information is also available for Brazil on a state-by-state basis in resolutions of 1x1 and 5x5 km.

GEOQUAL

ELC mapas

ECOGEO

Representa

DIV mapas

ColNucleo

FIGS\_R

*With the support and the collaboration of:*



**International Treaty on Plant Genetic Resources  
for Food and Agriculture**

Phone: (+39) 06 5705 6343 • Fax (+39) 06 570 56347 • [capfitogen@fao.org](mailto:capfitogen@fao.org)

<http://www.planttreaty.org/capfitogen>

