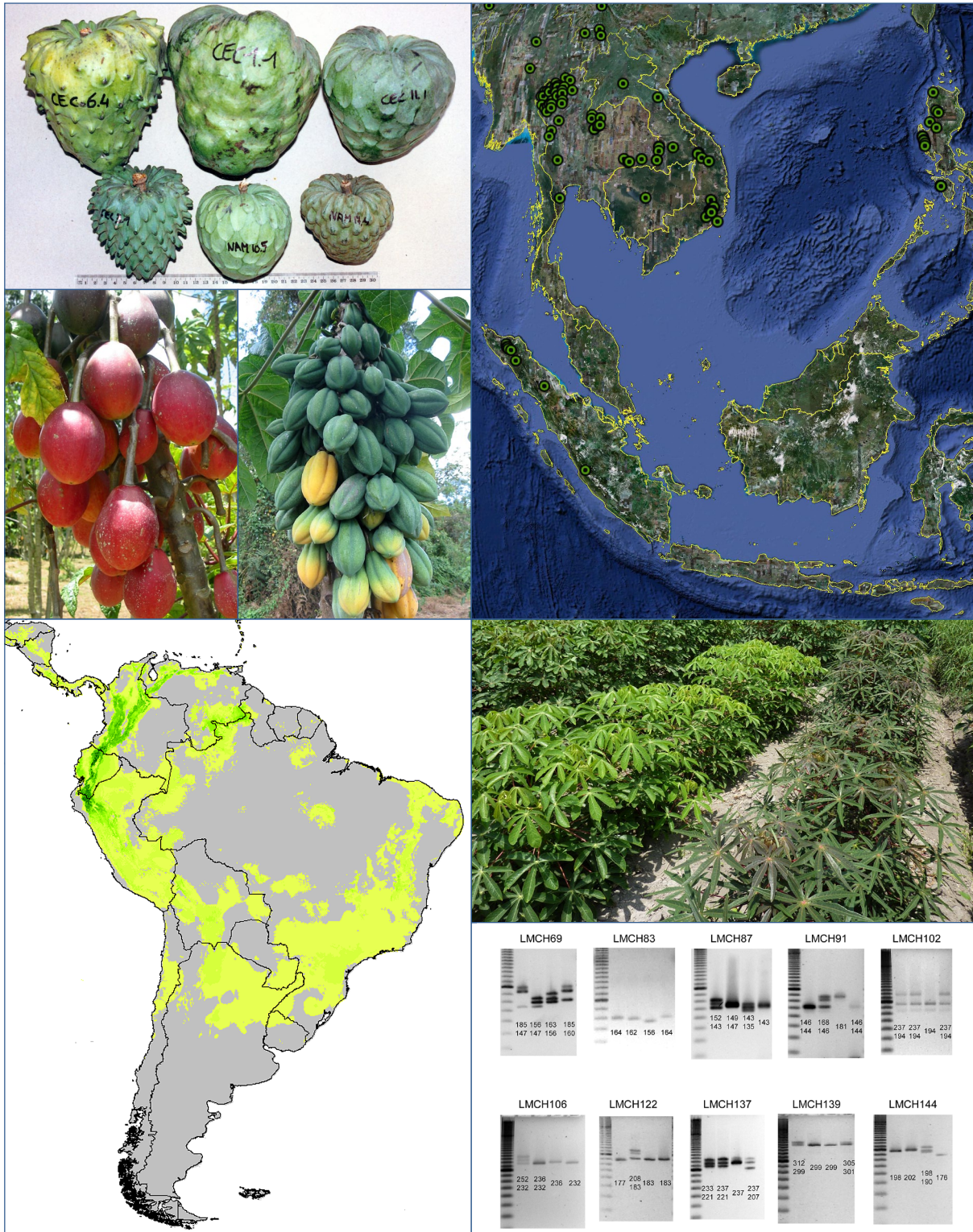


Training Manual on Spatial Analysis of Plant Diversity and Distribution

Xavier Scheldeman and Maarten van Zonneveld



Bioversity International is part of the Consultative Group on International Agricultural Research, which works to reduce hunger, poverty and environmental degradation in developing countries by generating and sharing relevant agricultural knowledge, technologies and policies. This research, focused on development, is conducted by a Consortium of 15 CGIAR centres working with hundreds of partners worldwide and supported by a multi-donor Fund.

The international status of Bioversity is conferred under an Establishment Agreement which, by December 2009, had been signed by the Governments of: Algeria, Australia, Belgium, Benin, Bolivia, Brazil, Burkina Faso, Burundi, Cameroon, Chile, China, Congo, Costa Rica, Côte d'Ivoire, Cyprus, Cuba, Czech Republic, Denmark, Ecuador, Egypt, Ethiopia, Ghana, Greece, Guinea, Hungary, India, Indonesia, Iran, Israel, Italy, Jordan, Kenya, Malaysia, Mali, Mauritania, Mauritius, Morocco, Norway, Oman, Pakistan, Panama, Peru, Poland, Portugal, Romania, Russia, Senegal, Slovakia, Sudan, Switzerland, Syria, Tunisia, Turkey, Uganda and Ukraine.

Financial support for Bioversity's research is provided by more than 150 donors, including governments, private foundations and international organizations. For details of donors and research activities please see Bioversity's Annual Reports, which are available in printed form on request from bioversity-publications@cgiar.org or from Bioversity's Web site (www.bioversityinternational.org).

The geographical designations employed and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of Bioversity or the CGIAR concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries. Similarly, the views expressed are those of the authors and do not necessarily reflect the views of these organizations.

Mention of a proprietary name does not constitute endorsement of the product and is given only for information.

Bioversity International
Via dei Tre Denari, 472/a,
00057 Maccarese, Rome, Italy
Tel.: (39) 0661181

bioversity-publications@cgiar.org

www.bioversityinternational.org

Bioversity International is the operating name of the International Plant Genetic Resources Institute (IPGRI) and the International Network for the Improvement of Banana and Plantain (INIBAP).

Supported by the CGIAR

Citation: Scheldeman, Xavier and van Zonneveld, Maarten. 2010. Training Manual on Spatial Analysis of Plant Diversity and Distribution. Bioversity International, Rome, Italy.

ISBN 978-92-9043-880-9

Cover: Credits: Photographs: Xavier Scheldeman (Bioversity International) - Maps: Maarten van Zonneveld (Bioversity International) - DNA gel electrophoresis images: Iñaki Hormaza (IHSM la Mayora, Consejo Superior de Investigaciones Científicas, Spain)

© Bioversity International, 2010

Contents

ACKNOWLEDGEMENTS	5
INTRODUCTION	6

SECTION A BASIC ELEMENTS AND DATA PREPARATION

1	INSTALLATION OF SOFTWARE AND EXAMPLE DATA FOR ANALYSIS	11
	1.1. INSTALLATION OF DIVA-GIS	11
	1.1.1. How to install DIVA-GIS	12
	1.2. INSTALLATION OF MAXENT	16
	1.2.1. How to install Java	16
	1.2.2. How to install Maxent	17
	1.3. INSTALLATION OF GOOGLE EARTH	18
	1.3.1. How to install Google Earth	18
	1.4. DATA FOR ANALYSIS	20
2	PREPARING AND IMPORTING DATA TO DIVA-GIS AND MAXENT	21
	2.1. PREPARING AND IMPORTING PRESENCE POINTS	23
	2.1.1. How to convert DMS data into DD format	25
	2.1.2. How to import georeferenced presence points to DIVA-GIS	28
	2.1.3. How to import georeferenced presence points in Maxent	29
	2.2. IMPORTING CLIMATE DATA TO DIVA-GIS AND MAXENT	34
	2.2.1. How to import climate data to DIVA-GIS	35
	2.2.2. How to import climate data in Maxent	36
	2.3. SOURCES OF SPATIAL AND OTHER RELEVANT DATA	38
3	BASIC ELEMENTS OF SPATIAL ANALYSIS IN DIVA-GIS	41
	3.1. VISUALIZATION IN DIVA-GIS	41
	3.1.1. How to perform basic visualizations using vector files	43
	3.1.2. How to perform basic visualizations using rasters	50
	3.1.3. How to combine rasters	57
	3.1.4. How to extract values from rasters based on presence points data	60
	3.1.5. How to create custom-made climate layers	62

3.1.6.	How to import generic climate data to DIVA-GIS	66
3.1.7.	How to make CLM files in DIVA-GIS	69
3.2.	EXPORTING LAYERS TO GOOGLE EARTH	74
3.2.1.	How to export data to Google Earth	74
3.3.	EDITING MAPS AND FINALIZING A PROJECT	77
4	QUALITY CONTROL	79
4.1.	QUALITY CONTROL BASED ON ADMINISTRATIVE UNIT INFORMATION	80
4.1.1.	How to verify data quality based on passport administrative unit data	80
4.2.	QUALITY CONTROL THROUGH THE IDENTIFICATION OF ATYPICAL POINTS	87
4.2.1.	How to identify outliers based on environmental data	88
SECTION B		
DATA ANALYSIS		
<hr/>		
5	SPATIAL ANALYSIS OF DIVERSITY FOR CONSERVATION PLANNING	97
5.1.	SPECIES RICHNESS	99
5.1.1.	How to carry out a spatial analysis of species richness	99
5.2.	INTRA-SPECIFIC DIVERSITY ANALYSIS BASED ON PHENOTYPIC DATA	111
5.2.1.	How to carry out a spatial diversity analysis using phenotypic data	112
5.3.	INTRA-SPECIFIC DIVERSITY ANALYSIS BASED ON MOLECULAR MARKER DATA	118
5.3.1.	How to carry out a spatial diversity analysis using molecular marker data	119
5.4.	IMPLICATIONS FOR THE FORMULATION OF CONSERVATION STRATEGIES	132
5.4.1.	How to identify priority zones for <i>in situ</i> conservation or germplasm collection	133
6	SPECIES DISTRIBUTION MODELLING AND ANALYSIS	139
6.1.	ANALYSIS OF THE REALIZED NICHE OF A SPECIES	140
6.1.1.	How to analyze and compare realized niches of different species	141
6.2.	MODELLING THE POTENTIAL DISTRIBUTION OF A SPECIES	147
6.2.1.	How to model the potential natural distribution of a plant species	148
6.3.	MODELLING THE IMPACT OF CLIMATE CHANGE ON SPECIES' DISTRIBUTION	159
6.3.1.	How to evaluate the impact of climate change on the distribution of species	159
6.4.	IDENTIFICATION OF GAPS IN COLLECTIONS OF WILD PLANT SPECIES	167
6.4.1.	How to identify possible gaps in collections	167

Acknowledgements

The authors are indebted to many colleagues and peers for their important role in the development of this manual.

We would like to acknowledge Nora Castañeda (Bioversity/CIAT) who, with her knowledge on using GIS methodologies to carry out spatial analysis of biodiversity data, participated in the initial phase of development. We also would like to give special thanks to Colm Bowe (Centre for Underutilized Crops, UK) whose feedback on the first versions of the manual contributed to shaping its current form, and to Jesús Salcedo for his most valuable support in the nitty gritty work of shaping the manual.

We are especially grateful to the “DIVA-GIS pioneers”, Robert Hijmans (University of California Davis, USA), Luigi Guarino (Global Crop Diversity Trust, Italy) and Andy Jarvis (CIAT, Colombia), for their constructive feedback. Their support and encouragement strongly motivated the authors to move forward with the manual’s preparation.

Many other persons also provided feedback on the various drafts of the manual. We would like to express our gratitude to our Bioversity colleagues, Karen Amaya, Margarita Baena, Michele Bozzano, Gea Galluzzi, Prem Mathur, Victoria Rengifo, Evert Thomas, Imke Thormann, Veerle Van Damme and Barbara Vinceti, as well as Sixto Iman (INIA Peru) and Diana Lara (CONIF, Colombia).

We would also like to thank Nicole Hoagland who, with her invaluable editorial assistance, considerably improved the manual’s readability.

The examples used in this manual have been tested during the delivery of various training courses. We are grateful to the organizers and donors of those courses organized between March 2006 and November 2010 in Argentina (Bariloche and Pergamino), Chile (Pucón), Colombia (Cali and Cartagena), Costa Rica (Turrialba), Ethiopia (Addis Abeba), Italy (Rome), Mali (Bamako) and Mexico (Mérida). We would like to particularly thank the course participants, as their feedback has contributed significantly to the improvement of the manual’s content.

Finally, the production of this manual would never have been possible without the financial support of INIA Spain (through the project “Strengthening Regional Collaboration in Conservation and Sustainable Use of Forest Genetic Resources in Latin America and Sub-Saharan Africa”) and of the Austrian Development Cooperation (through the project “Developing training capacity and human resources for the management of forest biodiversity”). We would like to acknowledge Ricardo Alia and Santiago Martínez of INIA for their positive and encouraging comments throughout the development process.

Introduction

Plant diversity is vital for the survival and well-being of humanity. A number of domesticated plant species are critical to global food security, while other species are of great importance for purposes such as wood and biofuel production. In addition to the cultivated species, many wild plants still play an important role in meeting local needs for food, fuel, medicine and construction materials; crop wild relatives are also of special interest for crop breeding programmes. There are currently hundreds of underutilized plant species and varieties displaying traits of interest to meet present and future needs, while the value of many other plant species is yet to be discovered.

The Convention on Biological Diversity (CBD), established in 1992, calls for a global strategy for plant conservation (CBD 2009a). In addition to the efforts of the CBD, the Global Plan of Action for the Conservation and Sustainable Utilization of Plant Genetic Resources for Food and Agriculture (GPA), adopted in 1996, and the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) (FAO 2009), entered into force in 2004, were formulated to focus on the potential of agricultural biodiversity and its importance for agricultural production. These international frameworks aim to increase information and actions to enhance the conservation and use of plant diversity. For example, Article 7 of the CBD calls for the identification and monitoring of biodiversity, paying particular attention to those species and varieties offering the greatest potential for sustainable use and requiring urgent conservation measures (CBD 2009b). Similarly, within the GPA, Priority Action 1 calls for increased surveying and inventorying of plant genetic resources for food and agriculture, while Priority Action 4 aims at promoting in situ conservation of crop wild relatives and wild plants for food production. Further, Priority Action 7 recommends planned and targeted collecting of plant genetic resources for food and agriculture. The importance of these activities is further confirmed in Article 5 of the ITPGRFA. In addition to such priorities, each of these international frameworks emphasizes the need to strengthen local capacities to carry out research related to diversity and genetic resources.

Spatial analysis can contribute significantly to the call for the improved understanding and monitoring of biodiversity. Results obtained from spatial analysis allow the formulation and implementation of more targeted, and hence more effective, conservation strategies. Outputs from spatial studies can provide critical information on the diversity present in specific geographic areas and can be used for various purposes; for example, to evaluate the current conservation status of plant species and to prioritise areas for conservation. Spatial information, combined with available characterization and/or evaluation data, has also proven useful for effective genebank management (e.g. definition of core collections, identification of collection gaps, etc.). This type of analysis is conducted using Geographic Information System (GIS) tools (Guarino et al. 2002), which allow one to carry out complex analyses combining different (spatial) data sources and generate clear maps, facilitating the uptake of outcomes by responsible authorities and encouraging the development and implementation of conservation policies. In recent years, technological advances and the growing availability of computers and GPS (Global Positioning System) receivers have led to the increased application of GIS analysis. The general accessibility and use of the internet has also created a revolution in the sharing of biodiversity, geographical and environmental data. The Global Biodiversity Information Facility (GBIF) is a platform providing public access to biodiversity data from national museums, herbaria and genebanks worldwide. In October 2010, the GBIF contained roughly 39 million georeferenced plant observations (GBIF 2009).

This training manual is intended for scientists (professionals and students) who work with biodiversity data and are interested in developing skills to effectively use spatial analysis programmes with GIS applications. It has been designed to serve as a self-teaching manual, but may also be used for training courses. The manual explains basic diversity and ecological analyses based on GIS applications. Results of these analyses offer a better understanding of spatial patterns of plant diversity, helping to improve conservation efforts. The training manual focuses on plants of interest for improving livelihoods (e.g. crops or crop wild relatives) and/or those which are endangered. Inter-specific and intra-specific diversity analyses using different types of data are presented: species presence, morphological characterization data (phenotypic data) and molecular marker data (data of DNA base pair compositions or molecular weights). Although this training manual focuses on plant diversity, many of the analyses described can also be applied when studying other organisms such as animals and fungi.

This manual has been published as a result of the increasing number of requests received by Bioversity International for capacity building on the spatial analysis of biodiversity data. The authors have developed a set of step-by-step instructions, accompanied by a series of analyses, based on free and publically available software: DIVA-GIS, a GIS programme specifically designed to undertake spatial diversity analysis; and Maxent, a species distribution modelling programme. The manual does not aim to illustrate the use of each individual DIVA-GIS and Maxent command/option, but focuses on using GIS tools to help answer common questions relating to the spatial analysis of biodiversity data. Throughout the manual, the importance of proper sampling is stressed; however, it is beyond the scope of the document to elaborate on sampling theories. The manual also does not discuss the statistical analysis of diversity data in detail; instead, when statistical methods and programmes are mentioned in the text, the reader is referred to alternative reference materials for further information.

After following the instructions and completing the analyses outlined in this manual, it is anticipated that the reader will have the capacity to carry out basic spatial diversity analyses to address common questions in conservation biology and plant genetic resources research.

References

- CDB. 2009a. Convention on Biological Diversity. Global Strategy for Plant Conservation [online]. Available from: <http://www.cbd.int/gspc>. Data accessed: October 2010.
- CDB. 2009b. Convention on Biological Diversity. Article 7. Identification and Monitoring [online]. Available from: <http://www.cbd.int/convention/articles.shtml?a=cbd-07>. Data accessed: October 2010.
- FAO. 2009. The International Treaty on Plant Genetic Resources for Food and Agriculture [online]. Available from: <http://www.planttreaty.org>. Data accessed: October 2010.
- GBIF. 2009. The Global Biodiversity Information Facility [online]. Available from: <http://www.gbif.org>. Data accessed: October 2010.
- Guarino L, Jarvis A, Hijmans RJ, Maxted N. 2002. Geographic Information Systems (GIS) and the conservation and use of plant genetic resources. In: Engels JMM, Ramanatha Rao V, Brown AHD, Jacson MT, editors. Managing plant genetic diversity. International Plant Genetic Resources Institute (IPGRI) Rome, Italy. pp. 387-404.



Section A

Basic elements and data preparation

Chapter 1

Installation of software and example data for analysis

In order to follow the instructions for each analysis presented in this manual, you will require the following computer programmes: DIVA-GIS, Maxent, Google Earth and Excel. The first three programmes are available online and free of charge at:

- <http://www.diva-gis.org>
- <http://www.cs.princeton.edu/~schapire/maxent>
- <http://earth.google.com>.

The Excel programme, however, is not freely available but is part of Microsoft Office, which is installed on most computers. If you do not have access to Excel, you may use Calc instead, a free software programme available through the OpenOffice package which can be downloaded at: <http://www.openoffice.org/>.

This manual refers to the most recent versions of the above-mentioned programmes, according to the date of the manual's publication. It is expected that the examples and subsequent step-by-step instructions included in this text will remain valid in the near future; however, it is acknowledged that these may change as new software versions are developed or if these programmes become obsolete. Therefore, it is recommended that the software and programme websites be consulted periodically for any possible updates which may be required.

The following paragraphs provide detailed instructions for downloading and installing DIVA-GIS and Maxent on your computer.

1.1. Installation of DIVA-GIS

DIVA-GIS is the main programme used to conduct and display the results of the spatial analyses presented in this manual. The installation of this programme is rather straightforward. All that is required is Version 7.3.0 (*diva730.zip*) of the software installer, available at: <http://www.diva-gis.org/download>.

Note

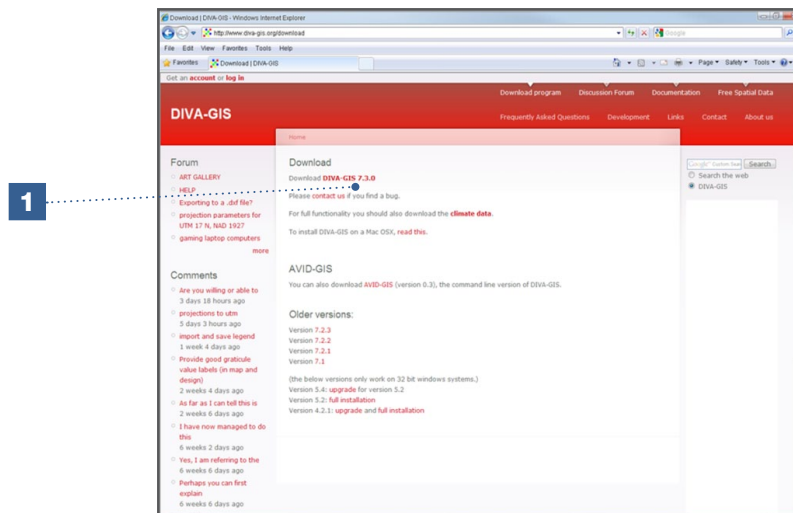
DIVA-GIS was developed for the Windows operating system, but it can also run on Mac OSX (see the DIVA-GIS website for additional information).

The DIVA-GIS website also offers climate data (see Section 2.3), thematic layers, a user manual, additional training materials and links to other relevant publications and websites.

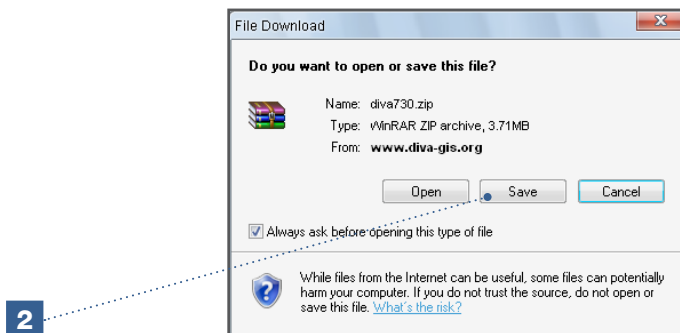
1.1.1. How to install DIVA-GIS

Steps:

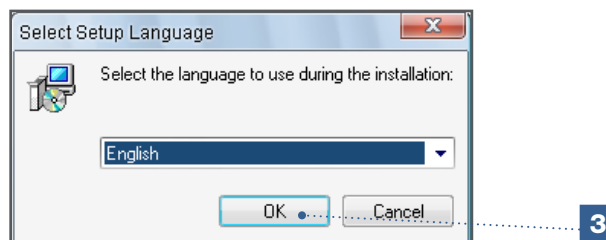
1. Download the compressed installer from the following URL: <http://www.diva-gis.org/download>.



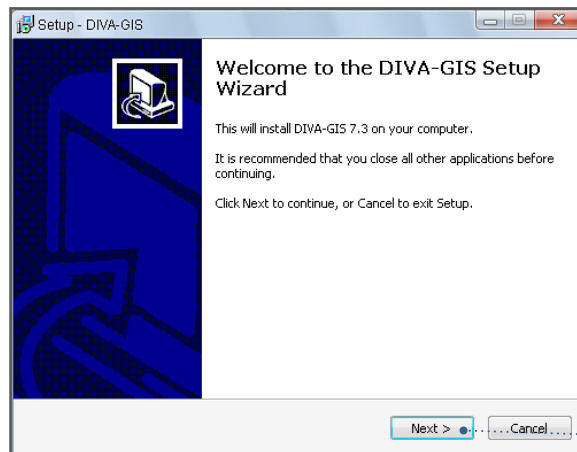
2. Save the file to your hard disk.



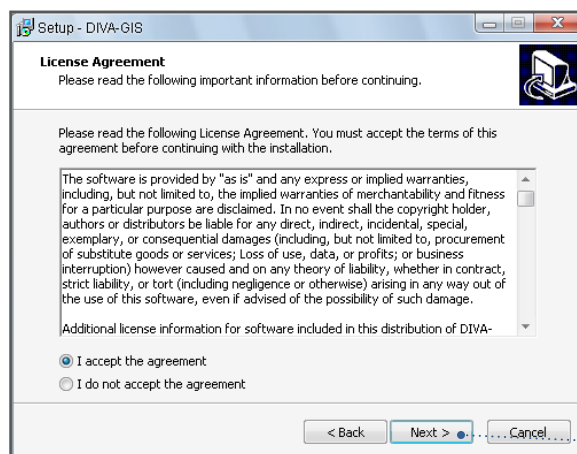
3. To execute the DIVA-GIS installer, first decompress (i.e. un-zip) the file saved in Step 2 above. To decompress the file, use free software such as 7-zip (<http://www.7-zip.org>) or IZARC (<http://www.izarc.org>). After decompressing the file, you will be able to view the *setup.exe* file. Click on *setup.exe*. The *Select Setup Language* window will then be displayed; select the language for the installation process. The DIVA-GIS programme itself is only available in English.



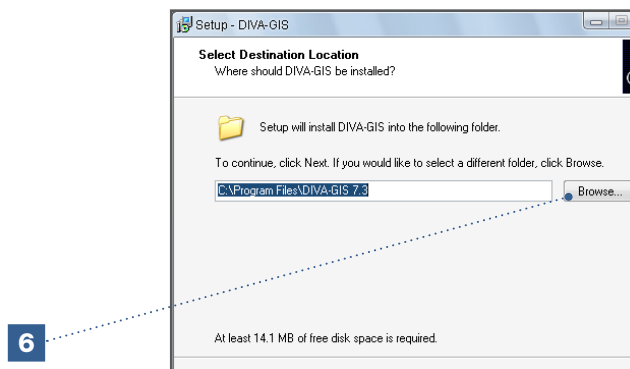
4. The Welcome message will be displayed. Click *Next* to continue.



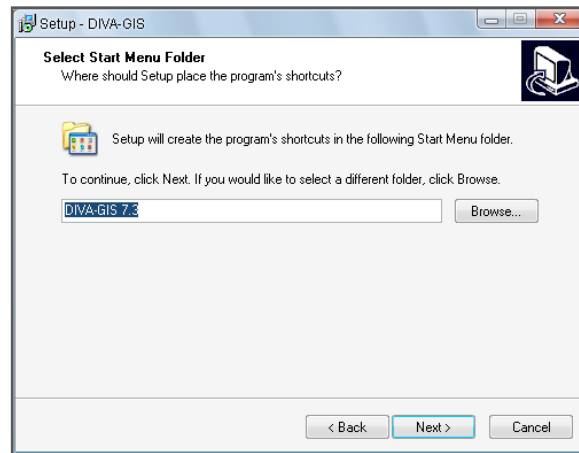
5. Carefully read the Terms of Agreement for the DIVA-GIS software license. If you agree, select the box: *I accept the agreement*. To continue, click *Next*. If you do not agree with the Licence Agreement you will not be able to install DIVA-GIS and the installation will be aborted.



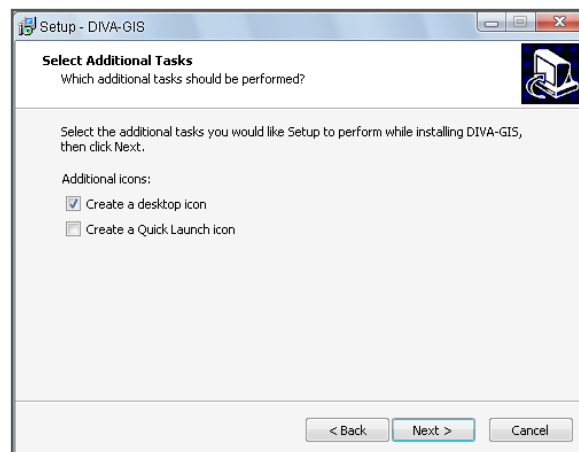
6. Indicate the file path where you would like to save the programme on your hard disk. The best option may be to use the default location automatically offered by the installer. If you prefer to install the programme in another location, select the file path using the *Browse* button.



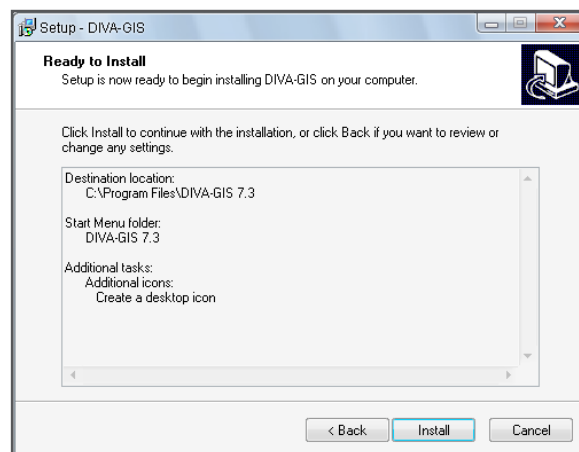
7. Indicate the folder from which you will have direct access to the programme in the *Startup Menu*. We recommend using the default option.



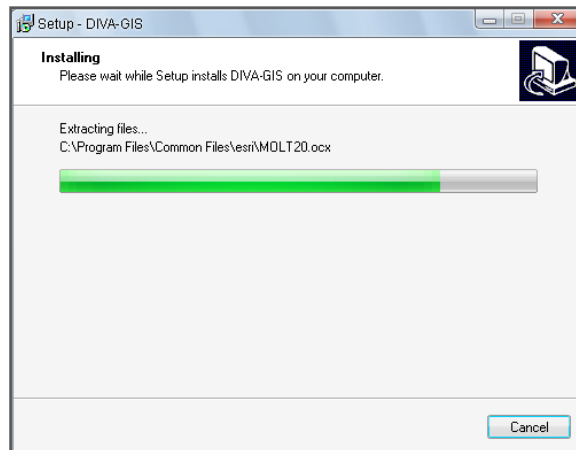
8. Select the desired configuration for the DIVA-GIS quick-access icons.



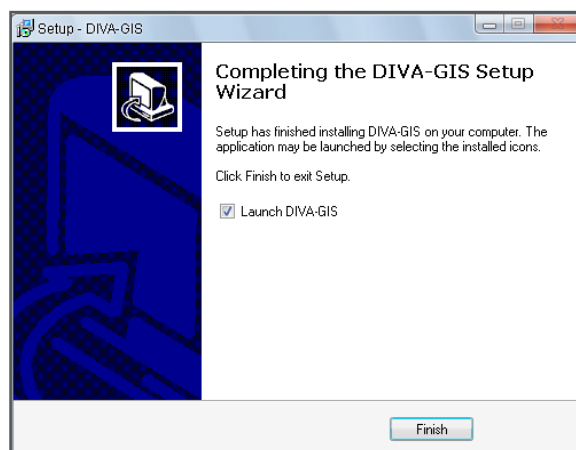
9. A window will be displayed summarizing the file path and options you selected. Review the information to ensure it is correct and proceed by clicking *Install*.



10. Installation will begin and the required files will be installed on your computer.



11. If all steps are completed correctly, the installation process should run smoothly. To finalize, click *Finish*.



1.2. Installation of Maxent

The Maxent programme will be used for the species distribution modelling analysis outlined in Chapter 6 of this manual. To download and use Maxent correctly, you will need to have Java installed in your computer. Most modern computers have Java installed by default, but it is also freely available at: <http://www.java.com>. If you find that Maxent will not run on your computer, it is most likely that you do not have Java installed.



1.2.1. How to install Java

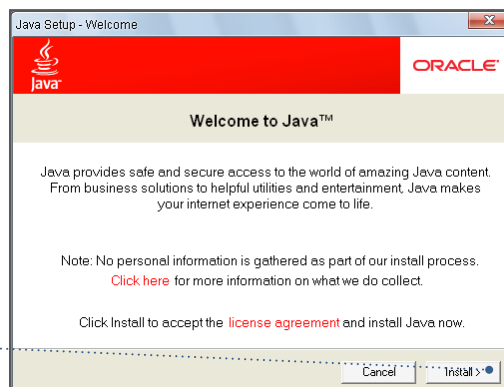
Steps:

1. If your navigator is Internet Explorer, go to: http://www.java.com/es/download/help/ie_online_install.xml.

If your navigator is Firefox, go to: http://www.java.com/en/download/help/firefox_online_install.xml.

2. Click *Install* to start the process.

2

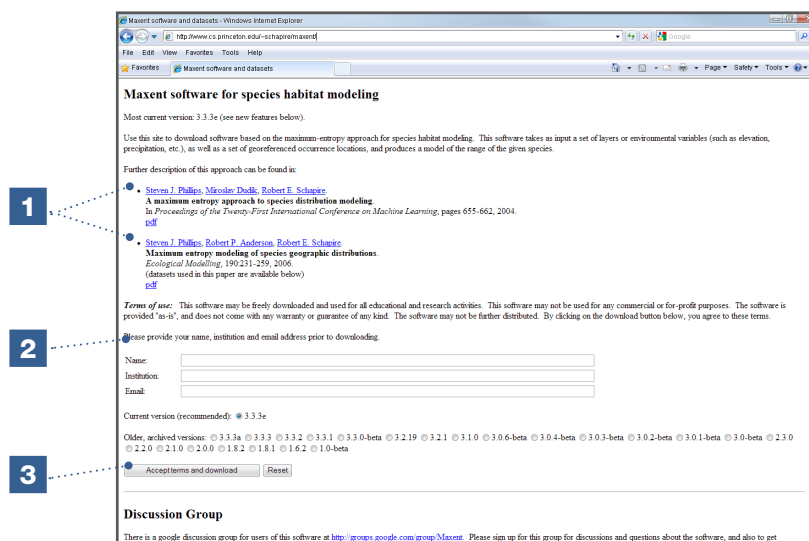


1.2.2. How to install Maxent

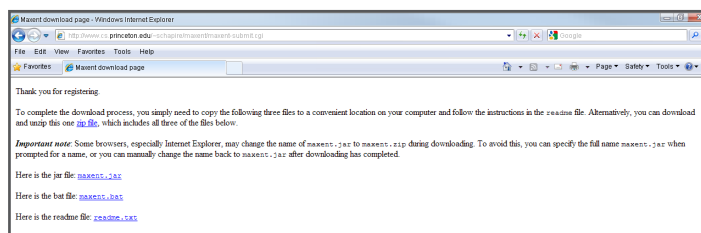
Maxent is an open-source programme which can be downloaded at: <http://www.cs.princeton.edu/~schapire/maxent/>. The programme's executable file is downloaded directly, so there is no need for separate installation.

Steps:

1. The documents explaining the concepts used by Maxent can be found on the website listed above.
2. Before downloading the programme, provide the requested contact information.
3. Finally, accept the license agreement and download the programme.



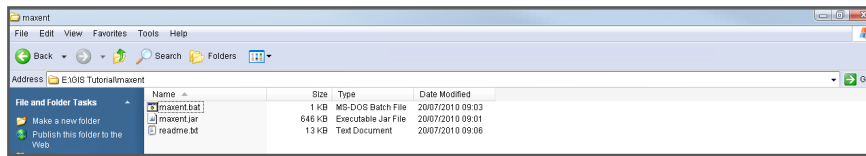
4. Maxent programme files can be downloaded separately or together in a compressed file format (*maxent.zip*).



5. Save the files in a convenient place on your hard disk. To start Maxent, double-click on the MS-DOS batch file (*.bat).

Note

If you have Winrar data compression software installed on your computer, it may be associated with JAR files, hindering the running of Maxent. In order to allow JAVA/Maxent to run, it might be necessary to remove the file association between Winrar and JAR files. To do this, unselect the JAR option on the 'Integration Tab' of the *Options/Settings* menu.



- If you run Maxent regularly, it may be useful to create a shortcut to the MS-DOS batch file (*.bat) on your desktop for quick access.

1.3. Installation of Google Earth

Google Earth has recently become an important tool for visualizing and sharing geographic information. Google Earth allows georeferenced data on the distribution of a taxon to be visualized in combination with high-quality satellite images. The increasing availability of these images will undoubtedly boost the use of Google Earth, while similar applications are also likely to appear in the market. With Google Earth, layers can be visualized in different scales ranging from the global level (continent, country, etc.) to levels as specific as a single tree (which could correspond to a presence point in a database), providing numerous visualization possibilities.

However, when using Google Earth, there is the risk of making high resolution interpretations with data of low precision. For example, you may know you have georeferenced species locations to a precision of 5 km, while the resolution of the Google Earth images may be as high as 1 x 1m. Trying to relate the species dataset to the Google image and its resolution would be meaningless in this instance.

1.3.1. How to install Google Earth

The installation of Google Earth is very straightforward and consists of downloading the installer, opening it and following the on-screen instructions. In order to be able to run Google Earth you must be connected to the internet (i.e. online).

Steps:

1. Download the Google Earth installer from <http://earth.google.com> (after accepting the conditions).



2. Open the downloaded installation file.
3. Follow the on-screen instructions. The installation will start by downloading additional data.

1.4. Data for analysis

In addition to installing the required software programmes, you will also need to download sample data to conduct the analyses outlined in the remaining sections of the manual.

For each section, a separate set of sample data is available; these datasets will be placed in one single folder (to facilitate access in DIVA-GIS). The datasets are available on Bioversity's website at: http://www.bioversityinternational.org/training/training_materials/GIS_manual.

Most analyses in this manual are based on detailed climate data (at a resolution of 2.5 minutes), meaning that the user must download large climate data files. This can be a constraint for users with poor internet connections.

- For those with a good internet connection, we suggest downloading the datasets in one zip file. Unzipping the file will create a new folder where all data will already be organized separately for use with the manual.
- For those with a slower internet connection, data can be downloaded section by section. Unzipping these different zip files will also organize the data as necessary for use with this manual.
- Should you have difficulties downloading the data, please do not hesitate to contact the manual's authors at bioversity-colombia@cgiar.org, x.schelde@gmail.com or m.vzonneveld@gmail.com. Data of lower resolution (climate data with less detail) can be provided or, if needed, the data can be sent in DVD format.

The datasets provided are based on existing studies but have been altered to improve their applicability for conducting analyses (e.g. errors have been introduced) or have been slightly modified to protect intellectual property (in the case of unpublished data). The datasets used in this manual should, by no means, be considered as appropriate for conducting specific studies. Section 2.3 provides sources of spatial data that can be used for this purpose.

Chapter 2

Preparing and importing data to DIVA-GIS and Maxent

The basis for spatial biodiversity analysis is observation data. Observation data are snapshots of species, trait or allele presence in time and space. To analyze observation data, users can use their own data, publically available data provided through international platforms such as the GBIF (<http://www.gbif.org>) or a combination of both. Much of the data in the GBIF are historical observations from herbaria and genebanks which may not reflect the current presence of taxa due to recent ecological processes or human interventions, such as forest conversion to agriculture and other changes in land use. Nonetheless, such data are still useful to gain insights into the ecological and genetic processes behind the geographical distribution of plant diversity. Observation data must be organized following the format specified by the applied GIS software. This chapter explains the type of data required for spatial analyses and species distribution modelling, and how to prepare and format data for use with DIVA-GIS and Maxent.

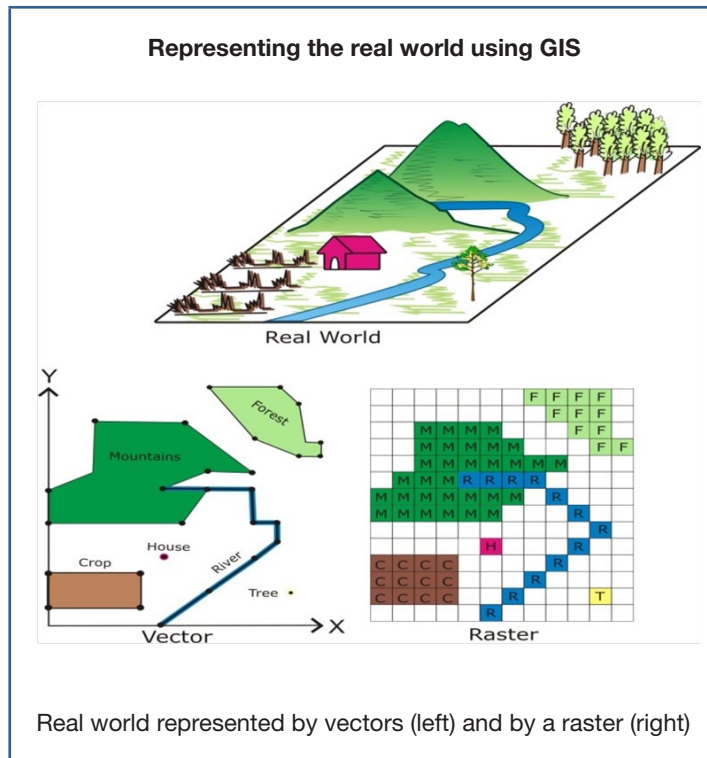
The analyses outlined in this manual are based on two components of observation data: passport data and associated data. Basic passport data consists of an identification code (ID), a taxonomic identification of an observed individual plant (or group of individuals) and the location of the collection or observation site. The analyses presented in this manual mainly use basic passport data, commonly referred to as 'presence points'. Presence points may be associated with additional data describing the collection or observation sites (e.g. land cover, soil type, etc.), the date of collection, source of coordinates, name(s) of collector(s) and the institution housing/conserving the specimen¹. Information about the collection date provides a time dimension to the analyses and can be used to analyze trends in species distribution (e.g. to monitor possible genetic erosion).

The opposite of presence points is absence data. While absence data can also be relevant in spatial analyses, e.g. monitoring trends in species distribution, it is often challenging to understand concrete reasons for the absence of taxa in a geographic unit, complicating the use of this data in ecological analyses (such as those that will be explained in Chapter 6). A taxon might be extinct due to human disturbance, or its absence might be explained by dispersal limitations or changes in the local environmental conditions. Further, the idea that plant collectors simply overlooked the presence of a specimen is always a possibility. Presence points, however, are generally much easier to relate to more credible factors such as environmental variables. Therefore, the analyses in this manual are based on presence points only.

Specimens are often associated with data providing further details as to individual plant characteristics and data used to assess intra-specific diversity and/or evaluation of agronomical traits, which can then be associated with spatial information (e.g. climate data). Thus, associated data can include information on morphological traits (e.g. size, shape, colours), physiology (e.g. days to germination, days to flowering), evaluation (e.g. yield, tolerance to abiotic and biotic stress) or DNA base pair composition (e.g.

¹ In crop genetic resources conservation such data are referred to as 'passport data'. For more information visit Bioversity's web page on Multi-crop Passport Descriptors (MCPD): [www.bioversityinternational.org/index.php?id=19&user_bioversitypublications_pi1\[showUid\]=2192](http://www.bioversityinternational.org/index.php?id=19&user_bioversitypublications_pi1[showUid]=2192).

molecular marker data). An analysis of diversity based on morphological characterization (phenotypic diversity) is outlined in Section 5.2, while the analysis presented in Section 5.3 is based on molecular marker data.



GIS software can process two different types of spatial data: vector and raster data. Vector data is composed of a set of georeferenced points which form either groups of points, lines or polygons and represent actual geographic areas. In this manual, the georeferenced points of plant individuals (presence points) or the administrative unit layers of countries (polygons) serve as examples of vector data. The most commonly used vector data file format is the shapefile.

DIVA-GIS uses shapefiles to represent vector data. A shapefile consists of (at least) three files with the same file name but a different file name extension. The extensions of these files are: '.shp', '.shx' and '.dbf'. Together, these three files make up a shapefile. The SHP file (*.shp) is the main file that stores spatial features and is the one shown in the dialogue box when opening the vector file using DIVA-GIS. The non-spatial attributes of the spatial features are stored in the dBase IV file (*.dbf), while the SHX file (*.shx) contains the indexing information of the vector file - it is used by computer programmes to more quickly access the *.shp file and link that file to the records in the dBase IV file (*.dbf).

Environmental data from specific geographic areas may also be georeferenced and are normally organized in a raster. Rasters consist of a grid of cells of identical size with each cell containing a value for a variable of interest (e.g. temperature, soil type). The optimal size of the cells depends on the size of the geographic area and on the objective of the study. The size of the study area is referred to as the 'extent', while the level of detail (given by the cell size) is called the 'resolution'. Using different cell sizes results in rasters with different resolutions (smaller cells generate rasters with higher resolution). Throughout this manual, different raster cell sizes are used, ranging from one (1) degree

(approximately 111 km at the equator) to 30 seconds (approximately 1 km at the equator)². Though still rather large geographic units, these sizes are appropriate for many types of spatial analyses and species distribution modelling at the national or regional level. These types of analyses are presented in Chapters 5 and 6.

Approximate size of geographic units (at the equator ²), rounded to km	
Degrees	Size
1 degree	111 km
10 minutes	18 km
5 minutes	9 km
2.5 minutes	5 km
30 seconds	1 km

GIS and species distribution modelling programmes use a range of different file types when processing raster data. The rasters used in the DIVA-GIS programme are in grid format, but the programme also allows one to import or export other types of raster file types (see Analyses 3.1.5 and 3.1.6).

A raster in DIVA-GIS consists of two files: one GRD file (*.grd) and one GRI file (*.gri); however, only the GRD file is shown when the raster is opened in DIVA-GIS. The GRD file contains the raster's general information, such as the position of raster corners, number of columns and rows and cell size; the GRI file contains the values for each raster cell and is therefore significantly larger.

When importing raster data to Maxent, it is recommended to use the ASCII II raster file type (which can be generated in DIVA-GIS from GRD files). An ASCII file consists only of one file (*.asc file). ESRI ArcGIS® software has its own file types, but can also import ASCII files.

2.1. Preparing and importing presence points

As mentioned previously, presence points can be compiled from data of vegetation inventories, plant collection expeditions or publicly available sources like the GBIF. Section 2.3 provides links to several information sources of this kind. GPS (Global Positioning System) equipment is now widely used to georeference presence points. Presence points can be organized in an Excel file and then converted into appropriate formats for spatial analysis using GIS programmes such as DIVA-GIS or species distribution modelling programmes such as Maxent. This section outlines the minimum required data for each presence point, how to format this data in Excel and how to then import such data to DIVA-GIS and Maxent.

² Towards the poles, metric distances in the longitudinal (east-west) direction will become shorter and become zero at the North and South Pole (where one can walk “around the world” in only a couple of meters). This manual will use cell sizes with degrees, which are easier to work with than cell sizes in metric distances.

To start preparing a database with presence points it is necessary to take into account:

a. Basic information

Presence points must include basic passport data of an individual plant or of a group of individuals in a specific geographic unit. They must include at least four elements: an identification code (ID), the taxonomic name of the individual plant (or the group of individuals) and longitude and latitude coordinates. These types of points are commonly used in spatial analysis of diversity and geographical distribution.

b. Storing coordinates

When geographic coordinates are used in DIVA-GIS, Maxent and other GIS and species distribution modelling programmes, it is preferable that they be reported using a latitude/longitude (lat-long) coordinate system and presented in Decimal Degrees (DD) format (DD.DDDD). When coordinates are available in Degrees, Minutes and Seconds (DMS) format (DD°MM'SS") or Degrees Minutes (DM) format (DD°MM.MM'), information should be converted to DD for use in a GIS or species distribution modelling programme. Therefore, the following formula is used for data conversion:

$$\text{Decimal degrees} = [(\text{Degrees } (^{\circ}) + \text{Minutes } (') / 60 + \text{Seconds } (") / 3600)] * H$$

H = 1 when the coordinate is in the Eastern (E) or Northern (N) Hemisphere

H = -1 when the coordinate is in the Western (W) or Southern (S) Hemisphere

Longitude	Degrees, Minutes & Seconds	Decimal Degrees	Latitude	Degrees, Minutes & Seconds	Decimal Degrees
Eastern Hemisphere	60°20'15" E	+ 60.3375	Northern Hemisphere	24°00'45" N	+ 24.0125
Western Hemisphere	60°20'15" W	- 60.3375	Southern Hemisphere	24°00'45" S	- 24.0125

It is recommended that DD points have a precision of at least four decimals. Conversion of DMS or DM data into DD format in Excel can be conducted by using the text functions (DMS data are stored as text) RIGHT/MID/LEFT.

Note

Be careful not to generate a false sense of precision by creating more decimals when converting data from DMS to DD format. A coordinate in DMS format including information on only the degrees and minutes (e.g. 60°05') has an actual precision of two decimals, but can be presented in DD format as a coordinate with four decimals (60.0833). When presence point data includes coordinates in DMS format with less than four decimals (only degrees and minutes or only degrees) it is recommended to add an extra field to specify the precision of the coordinates, before converting these from DMS into DD format.

In addition to the latitude/longitude coordinate system, another common coordinate system is Universal Transverse Mercator (UTM). While DIVA-GIS can operate using UTM, the latitude/longitude system (in DD format) is still preferred as it is more likely to be compatible with the available thematic layers (administrative unit information, climate, land cover). UTM coordinates can be converted into the latitude/longitude system in decimal degrees by using Excel calculation sheets; these are available on the internet from sites such as: <http://www.uwgb.edu/dutchs/UsefulData/UTMFormulas.htm>.

The *Projection* option in the *Tools* menu of DIVA-GIS also allows you to convert lat-long formats into UTM or vice versa.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel • Maxent and Java 	<p>Data Files:</p> <p>Folder 2.1 Importing observation data</p> <ul style="list-style-type: none"> • <i>Vcundinamarcensis_DMSdata.xls</i> • <i>Vasconcellea.xls</i>

2.1.1. How to convert DMS data into DD format

This exercise uses the *Vcundinamarcensis_DMSdata.xls* Excel file as an example; the file can be found in the data folder accompanying this chapter and consists of seven spreadsheets in which DMS coordinates are calculated stepwise into DD.

Steps:

1. Open the *Vcundinamarcensis_DMSdata.xls* Excel file and select the *Start data* spreadsheet. Note that the latitude and longitude coordinates are presented in DMS.

ID	SPECIES	LATITUDE	LONGITUDE	COUNTRY	ADM1
3433	V. cundinamarcensis	6°57'48"N	75°25'03"W	Colombia	Antioquia
3037	V. cundinamarcensis	7°10'17"N	75°45'47"W	Colombia	Antioquia
2816	V. cundinamarcensis	7°01'59"N	75°18'59"W	Colombia	Antioquia
2836	V. cundinamarcensis	6°19'59"N	75°15'00"W	Colombia	Antioquia
3030	V. cundinamarcensis	6°54'00"N	75°37'51"W	Colombia	Antioquia
3418	V. cundinamarcensis	6°09'59"N	75°33'59"W	Colombia	Antioquia

2. Go to the *Columns* spreadsheet. Create separate columns for *degrees*, *minutes* and *seconds* of the latitude and longitude (Columns E, F, G and L, M, N). The *Lat* column (Column D) has been created to indicate whether the latitude coordinate is in the northern hemisphere, N (1), or in the southern hemisphere, S (-1). The *Lon* column (Column K) has been created to indicate whether the longitude coordinate is in the eastern hemisphere, E (1), or in the western hemisphere, W (-1).

ID	SPECIES	LATITUDE	Lat	degrees	minutes	seconds	decimal degree/latitude	LONGITUDE	Lon	degrees	minutes
3433	V. cundinamarcensis	6°57'48"N	1		6	57	48	75°25'03"W	-1	75	25
3037	V. cundinamarcensis	7°10'17"N						75°45'47"W			
2816	V. cundinamarcensis	7°01'59"N						75°18'59"W			
2836	V. cundinamarcensis	6°19'59"N						75°15'00"W			
3030	V. cundinamarcensis	6°54'00"N						75°37'59"W			
3418	V. cundinamarcensis	6°09'59"N						75°33'59"W			

- Open the *DD formula* spreadsheet. The formula to convert latitude and longitude from DMS to DD ($=+D2*(E2+(F2/60)+(G2/3600))$) has been inserted in the latitude and longitude columns named *decimal degrees* (Columns H and O).

ID	SPECIES	LATITUDE	Lat	degrees	minutes	seconds	decimal degrees	latitude	LONGITUDE	Lon	degrees	minutes
2	3433 V. cundinamarcensis	6°57'49"N	1	57	48	6.9633		75°25'03"W		-1	75	
3	3037 V. cundinamarcensis	7°01'17"N						75°45'47"W				
4	2816 V. cundinamarcensis	7°01'59"N						75°18'59"W				
5	2636 V. cundinamarcensis	6°19'59"N						75°15'00"W				
6	3030 V. cundinamarcensis	6°54'00"N						75°57'59"W				
7	3418 V. cundinamarcensis	6°09'59"N						75°33'59"W				

- Go to the *Formulas* spreadsheet. Special formulas (based on the Excel text functions RIGHT/MID/LEFT) are inserted in the columns named *Lat*, *Lon*, *degrees*, *minutes* and *seconds* (Columns D and K, Columns E, F, G and L, M, N) in order to separate the values of the hemisphere, degrees, minutes and seconds from the *Latitude* and *Longitude* columns (Columns C and J) into separate columns.
- Go to the *Copied formulas* spreadsheet. The formulas are copied in each row to determine the latitude and longitude in DD for every presence point.

ID	SPECIES	LATITUDE	Lat	degrees	minutes	seconds	decimal degrees	latitude	LONGITUDE	Lon	degrees	minutes
2	3433 V. cundinamarcensis	6°57'49"N	1	57	48	6.9633		75°25'03"W		-1	75	
3	3037 V. cundinamarcensis	7°01'17"N	1	17	10	7.1714		75°45'47"W		-1	75	
4	2816 V. cundinamarcensis	7°01'59"N	1	17	01	7.0331		75°18'59"W		-1	75	
5	2636 V. cundinamarcensis	6°19'59"N	1	19	59	6.3331		75°15'00"W		-1	75	
6	3030 V. cundinamarcensis	6°54'00"N	1	54	00	6.9000		75°57'59"W		-1	75	
7	3418 V. cundinamarcensis	6°09'59"N	1	09	59	6.1664		75°33'59"W		-1	75	

- Open the *Values* spreadsheet. Copy the calculated coordinates in DD (Columns H and O) and paste the values, using the *Paste special* option, into the respective *latitude* and *longitude* columns (Columns I and P). This is necessary as DIVA-GIS cannot import new DD values if they are still presented as formulas.
- Open the *Final* spreadsheet. The coordinates in DMS are converted to DD and the presence point database is ready to be imported to DIVA-GIS. It is important to keep the original DMS coordinates in order to track any errors that may have occurred during the calculation of coordinates.

ID	SPECIES	latitude	longitude	COUNTRY	ADM1	LATITUDE	LONGITUDE
2	3433 V. cundinamarcensis	6.9633	-75.4175	Colombia	Antioquia	6°57'48"N	75°25'03"W
3	3037 V. cundinamarcensis	7.1714	-75.7631	Colombia	Antioquia	7°01'17"N	75°45'47"W
4	2816 V. cundinamarcensis	7.0331	-75.3164	Colombia	Antioquia	7°01'59"N	75°18'59"W
5	2636 V. cundinamarcensis	6.3331	-75.2500	Colombia	Antioquia	6°19'59"N	75°15'00"W
6	3030 V. cundinamarcensis	6.9000	-75.9864	Colombia	Antioquia	6°54'00"N	75°57'59"W
7	3418 V. cundinamarcensis	6.1664	-75.5664	Colombia	Antioquia	6°09'59"N	75°33'59"W

Georeferencing presence information

Sometimes presence points lack geographic coordinates and have only a description of their location in the form of administrative unit data. These data can be classified as follows:

- Country
- Administrative unit level 1 (Adm1): state, department, region, province (of a country)
- Administrative unit level 2 (Adm2): province, canton, municipality
- Locality: city, town, national park, etc.

	A	B	C	D	E	F
1	ID	Taxon	Country	Adm1	Adm2	Locality
2	1	<i>Capsicum chinense</i> Jacq.	Bolivia	Pando	Nicolas Suarez	Cobjija
3	2	<i>Capsicum chinense</i> Jacq.	Bolivia	Pando	Nicolas Suarez	Cobjija
4	55	<i>Capsicum frutescens</i> L.	Bolivia	Beni	Vaca Diez	Riberalta
5	72	<i>Capsicum eximium</i> Hunz.	Bolivia	Tarija	Mendez	San Lorenzo
6	73	<i>Capsicum eximium</i> Hunz.	Bolivia	Tarija	Mendez	San Lorenzo
7						
8						

Example of administrative unit data with *Capsicum* accessions originating from Bolivia.

In such cases, databases known as ‘gazetteers’ can be referenced for assistance. Gazetteers are lists of administrative units (e.g. municipalities) with respective geographic coordinates that can assist in assigning georeferenced information to the points of interest. Gazetteers with administrative unit data for most countries are available online and are freely accessible from the DIVA-GIS web page: <http://www.diva-gis.org/gdata>. The files are generally in dBase IV format (*.dbf), which can be opened and searched in Excel (using the *Edit/Find* option).

However, it may not always be possible to georeference a site using a gazetteer. This might be the case if information on the locality is incomplete or simply not available. Another difficulty may arise if several places have the same name (see the table below). In this situation, differences at higher administrative levels, such as Adm1, can help to distinguish between places and resolve the conflict.

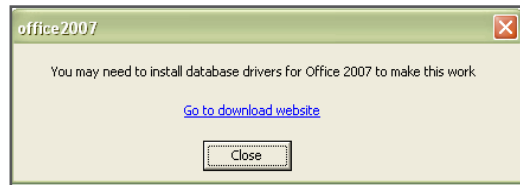
In addition to gazetteers, other programmes such as Biogeomancer, available at: <http://www.biogeomancer.org>, allow georeferencing of points based on site information.

	A	B	C	D	E	F	G	H	I
13106	Jalapa	P	PPL	24 8833	101 5500	Coahuila De Zaragoza			
13107	Jalapa	P	PPL	20 9866	101 7186	Guanajuato			
13108	Jalapa	P	PPL	20 1666	102 0166	Michoacan de Ocampo			
13109	Jalapa	P	PPL	19 5333	96 9186	Veracruz-Llave			
13110	Jalapa	P	PPL	17 7166	92 8186	Tabasco			
13111	Jalapa	P	PPL	17 6333	99 5666	Guerrero			
13112	Jalapa	P	PPL	17 3333	99 2666	Guerrero			
13113	Jalapa	P	PPL	16 5000	95 4666	Oaxaca			
13114	Jalapa	P	PPL	16 3500	92 6666	Chiapas			

Gazetteer for Mexico showing more than one site referred to as ‘Jalapa’.

2.1.2. How to import georeferenced presence points to DIVA-GIS

As mentioned, presence points can be imported from Excel to DIVA-GIS. An alert sign (shown below) will appear on the screen if you attempt to import data directly from Excel 1997-2003 to DIVA-GIS.



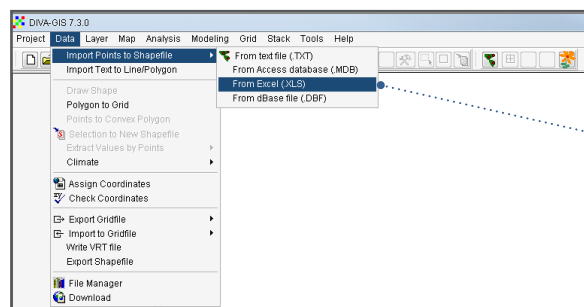
The alert sign indicates that you may need to install database drivers for Office 2007. Clicking on *Go to download website* directs you to the Microsoft Download Center to get the respective drivers.

Note

People who work with another program than Excel to manage their passport database or still have problems importing presence points from Excel to DIVA-GIS, after having installed the drivers, can import alternatively their points to DIVA-GIS from text files (*.txt), CSV files (*.csv) or dBase IV files (*.dbf). Importing points from these file types to DIVA-GIS is similar to introducing them from Excel following the steps explained in this example, except that the presence point database has to be saved as another file type. For further explanation about importing presence points from these file types please consult the DIVA-GIS Operating Manual, available online at: <http://www.diva-gis.org/documentation>.

Steps for importing data from Excel to DIVA-GIS:

1. Excel 2007: Save your presence points as an Excel file 1997-2003 (*.xls) and close Excel 2007. [If you save as Excel file (*.xlsx) the presence points cannot be imported to DIVA-GIS]. Earlier versions of Excel: Save your presence points as an Excel file (*.xls).
2. Open DIVA-GIS and go to *Data/Import Points/From Excel (.XLS)*

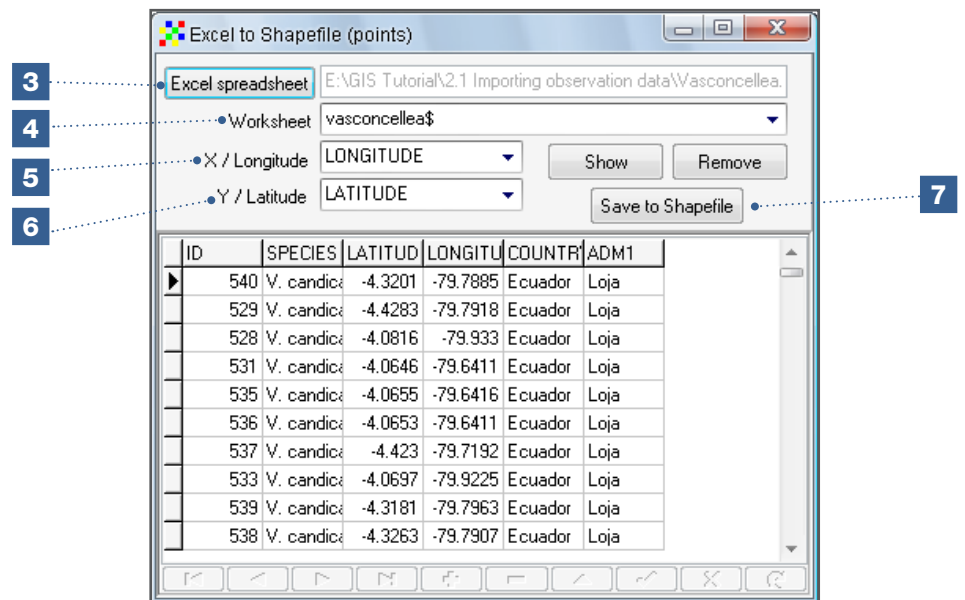


- In *Excel spreadsheet* select the Excel file that has the presence points you would like to import (*Vasconcellea.xls*) to DIVA-GIS.

Note

Always close the Excel file before importing data to a vector file in DIVA-GIS.

- Under *Worksheet*, select the Excel sheet with the data you wish to import into DIVA-GIS.
- Select the column with the longitude coordinates.
- Select the column with the latitude coordinates.
- Click on *Save to Shapefile* to generate a vector file (*.shp).



2.1.3. How to import georeferenced presence points in Maxent

Maxent is a species distribution modelling programme used for predicting the potential distribution of one or more species. As with DIVA-GIS, Maxent works with georeferenced presence data. This section illustrates how to import georeferenced presence points to Maxent. A detailed explanation of how Maxent works is presented in Section 6.2.

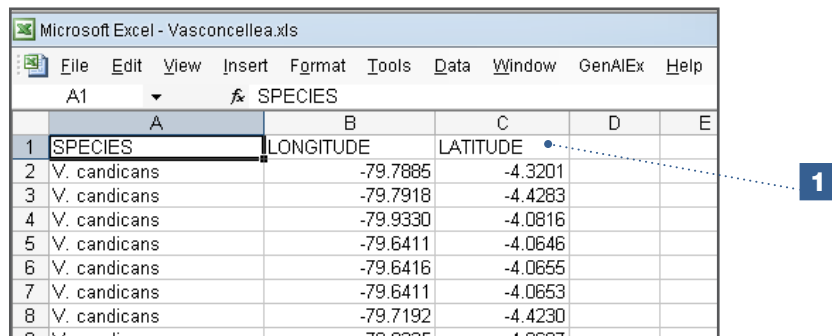
In Maxent, data must be imported as a Comma Separated Values (CSV)(Comma delimited) file (*.csv) and should include three fixed fields (columns) corresponding to the following categories: Species, Longitude and Latitude. Columns should be listed in this specific order (but additional columns with more information are permitted). The three-column file can be prepared by starting with a database originally constructed in Excel and then saved in CSV (Comma delimited) (*.csv) format. There are slight differences in procedure when creating CSV files using Excel 2007 and Excel 1997-2003.

Steps:

Steps for preparing a CSV (Comma delimited) file (*.csv) for Maxent in **Excel 1997-2003** (Hereafter the steps for preparing a CSV file (*.csv) in the Excel 2007 are explained).

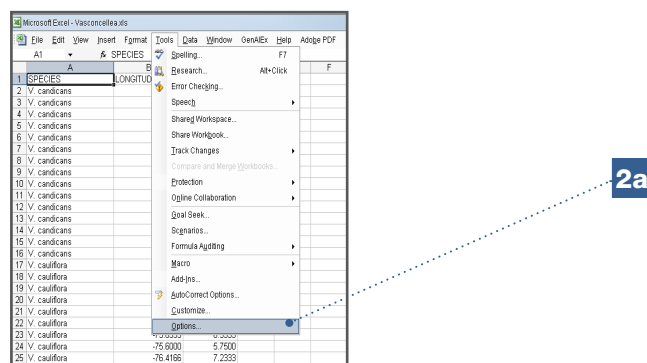
The Excel file, *Vasconcellea.xls*, is used in the following example; the file can be found in the data folder accompanying this section.

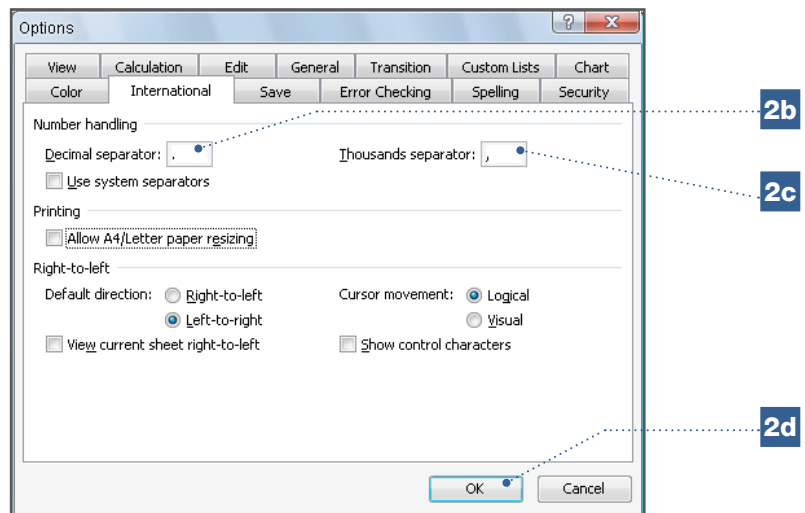
1. In a new Excel sheet, copy the data for the variables: Species, Longitude and Latitude, **in this specific order**.



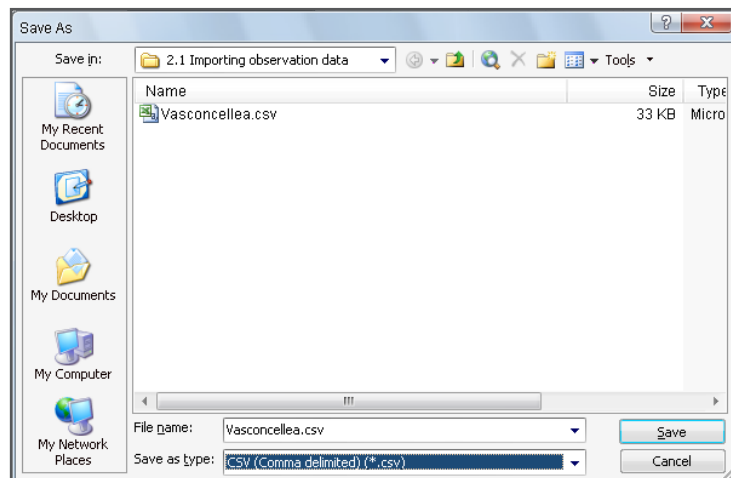
	A	B	C	D	E
1	SPECIES	LONGITUDE	LATITUDE		
2	V. candicans	-79.7885	-4.3201		
3	V. candicans	-79.7918	-4.4283		
4	V. candicans	-79.9330	-4.0816		
5	V. candicans	-79.6411	-4.0646		
6	V. candicans	-79.6416	-4.0655		
7	V. candicans	-79.6411	-4.0653		
8	V. candicans	-79.7192	-4.4230		

2. Before proceeding, make sure the decimals are effectively separated by points:
 - a. Go to Tools/ Options
 - b. Under the *International* tab, select *points* [.] to separate decimals
 - c. Under the *International* tab, select *commas* [,] to separate thousands
 - d. Click OK.





3. Make sure no commas are used in the characters in the column for species name(s) or in any additional columns, if present. This will help to avoid errors originating from unintentional separations in the information of the CSV file (*.csv), e.g. in the administrative unit information.
4. Save file as CSV file (*.csv).

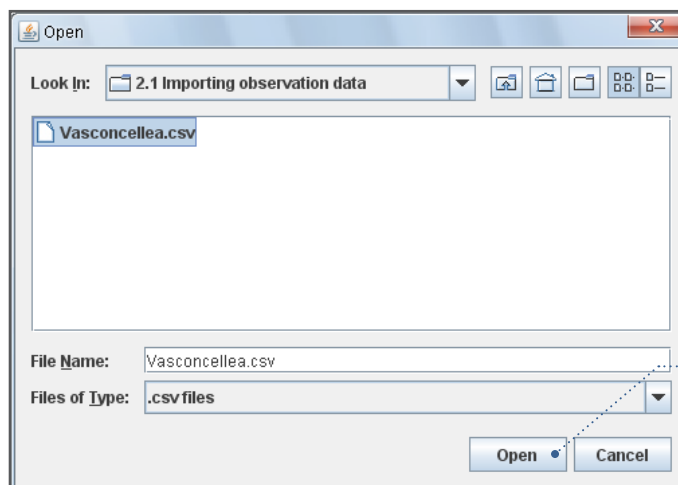
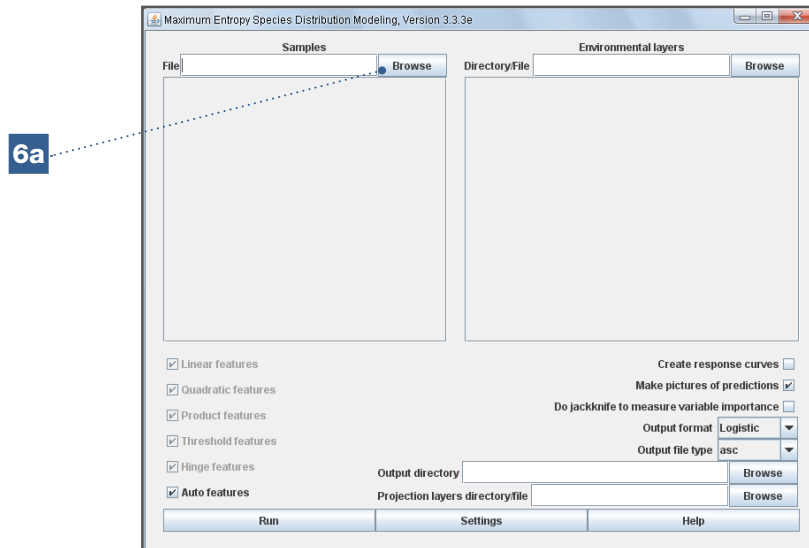


5. In Excel, open the CSV file (*.csv) to verify that coordinates have all the decimals required.

6. Open Maxent; under *Samples* indicate the CSV file (*.csv) with the georeferenced presence points.

To select the file:

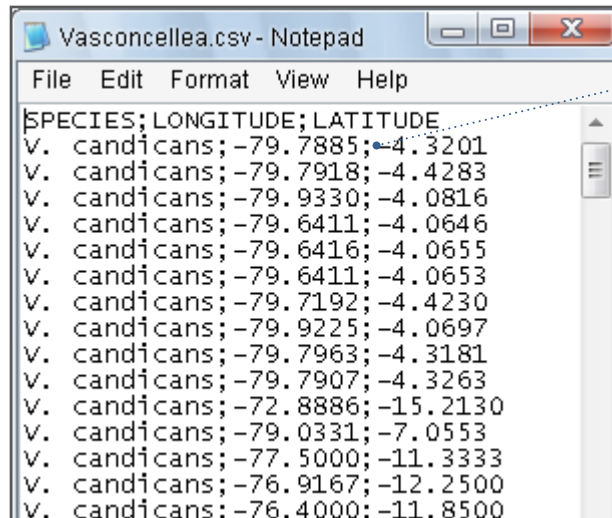
- a. Go to the *Browse* option
- b. Open the CSV file (*.csv).



Steps for preparing a Comma Separated Values (CSV) file for Maxent in Excel 2007:

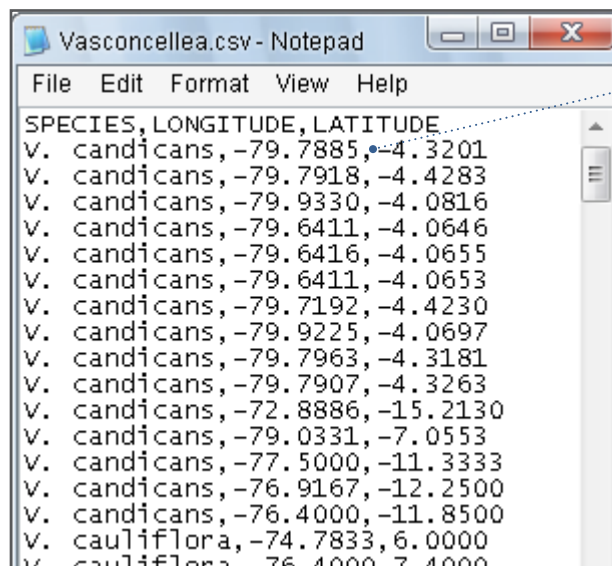
1. In a new sheet, copy the data for the variables: Species, Longitude and Latitude, **in this specific order**.
2. Check that decimals in Excel are separated by points;
 - a. Go to *Office Button/Excel Option*
 - b. Under the *Advanced* tab, select *points* [.] to separate decimals
 - c. Under the *Advanced* tab, select *commas* [,] to separate thousands.
3. Save the CSV file (*.csv) and open it using *Notepad*.

4. Make sure the different data are separated by commas; if this is not the case, make the necessary changes. This can be done using the *Replace* function in the *Edit/Replace* menu.
 - a. Delimited by semicolons
 - b. Replacement by commas
 - c. Remember to save the changes made to the file
 - d. Finally, follow Step 6 in the section above (for Excel 1997-2003) to open the CSV file (*.csv) file in Maxent.



```

Vasconcellea.csv - Notepad
File Edit Format View Help
SPECIES; LONGITUDE; LATITUDE
V. candicans; -79.7885; -4.3201
V. candicans; -79.7918; -4.4283
V. candicans; -79.9330; -4.0816
V. candicans; -79.6411; -4.0646
V. candicans; -79.6416; -4.0655
V. candicans; -79.6411; -4.0653
V. candicans; -79.7192; -4.4230
V. candicans; -79.9225; -4.0697
V. candicans; -79.7963; -4.3181
V. candicans; -79.7907; -4.3263
V. candicans; -72.8886; -15.2130
V. candicans; -79.0331; -7.0553
V. candicans; -77.5000; -11.3333
V. candicans; -76.9167; -12.2500
V. candicans; -76.4000; -11.8500
  
```



```

Vasconcellea.csv - Notepad
File Edit Format View Help
SPECIES, LONGITUDE, LATITUDE
V. candicans, -79.7885, -4.3201
V. candicans, -79.7918, -4.4283
V. candicans, -79.9330, -4.0816
V. candicans, -79.6411, -4.0646
V. candicans, -79.6416, -4.0655
V. candicans, -79.6411, -4.0653
V. candicans, -79.7192, -4.4230
V. candicans, -79.9225, -4.0697
V. candicans, -79.7963, -4.3181
V. candicans, -79.7907, -4.3263
V. candicans, -72.8886, -15.2130
V. candicans, -79.0331, -7.0553
V. candicans, -77.5000, -11.3333
V. candicans, -76.9167, -12.2500
V. candicans, -76.4000, -11.8500
V. cauliflora, -74.7833, 6.0000
V. cauliflora, -76.4000, 7.4000
  
```

2.2. Importing climate data to DIVA-GIS and Maxent

The *Bioclim/Domain* option in the *Modeling* menu in DIVA-GIS, allows one to carry out multiple analyses based on climate data. These analyses include the identification of atypical points known as outliers (see Chapter 4), the delineation of an ecological niche, the prediction of potential species distribution and the subsequent gap analysis or analysis of climate change impacts (see Chapter 6).

Before starting any of these analyses, you must import climate data to DIVA-GIS. Such data is freely available from global databases. The most commonly referenced database is Worldclim (Hijmans et al. 2005), which uses 19 derived bioclimatic variables (Busby 1991) in addition to monthly climate data (maximum and minimum temperature and precipitation). Compared to the commonly used temperature and precipitation parameters, the 19 bioclimatic variables are more directly related to the physiologic aspects of plant growth and do not consider the timing when a particular state occurs, i.e. it does not matter whether the hottest month is July (Northern hemisphere) or January (Southern hemisphere). Some bioclimatic variables include typical, basic climate parameters (e.g. BIO1, mean annual temperature or BIO12, annual precipitation), while others combine temperature and precipitation in one variable (e.g. BIO18, precipitation during warmest quarter). Others capture aspects of seasonality (e.g. BIO4 for temperature, BIO15 for precipitation), which can also be important to determine a species' distribution.

The 19 Bioclimatic Variables³

BIO1 = Annual mean temperature
 BIO2 = Mean diurnal range (max temp – min temp) (monthly average)
 BIO3 = Isothermality (BIO1/BIO7) * 100
 BIO4 = Temperature Seasonality (Coefficient of Variation)
 BIO5 = Max Temperature of Warmest Period
 BIO6 = Min Temperature of Coldest Period
 BIO7 = Temperature Annual Range (BIO5-BIO6)
 BIO8 = Mean Temperature of Wettest Quarter
 BIO9 = Mean Temperature of Driest Quarter
 BIO10 = Mean Temperature of Warmest Quarter
 BIO11 = Mean Temperature of Coldest Quarter
 BIO12 = Annual Precipitation
 BIO13 = Precipitation of Wettest Period
 BIO14 = Precipitation of Driest Period
 BIO15 = Precipitation Seasonality (Coefficient of Variation)
 BIO16 = Precipitation of Wettest Quarter
 BIO17 = Precipitation of Driest Quarter
 BIO18 = Precipitation of Warmest Quarter
 BIO19 = Precipitation of Coldest Quarter

³ Original list and further reading: <http://www.worldclim.org/bioclim>.

Climate data are imported to DIVA-GIS from CLM (*.clm) files, available on the DIVA-GIS website (<http://www.diva-gis.org/climate.htm>) or are prepared for a specific study area based on global climate data (see Analysis 3.1.6). From the CLM files, basic temperature parameters (maximum and minimum temperature and precipitation), as well as values for the 19 bioclimatic variables can be extracted in DIVA-GIS, as either layers or associated data with observation points. CLM files in different resolution are available at: <http://www.diva-gis.org/climate.htm>.

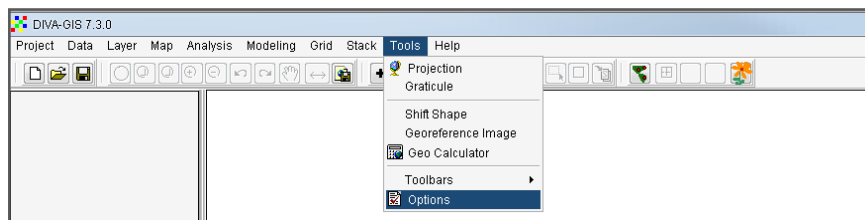
PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel • Maxent and Java 	<p>Data Files:</p> <p>Folder 2.2 Importing climate data</p> <ul style="list-style-type: none"> • Folder diva_worldclim_2-5min (zip file) downloaded from http://www.diva-gis.org/climate • Folder wclim_eth_2-5min_ascii (asc files)

2.2.1. How to import climate data to DIVA-GIS

This section explains how to import climate data to DIVA-GIS from CLM files (*.clm). DIVA-GIS uses the CLM format to store and read spatial climate data.

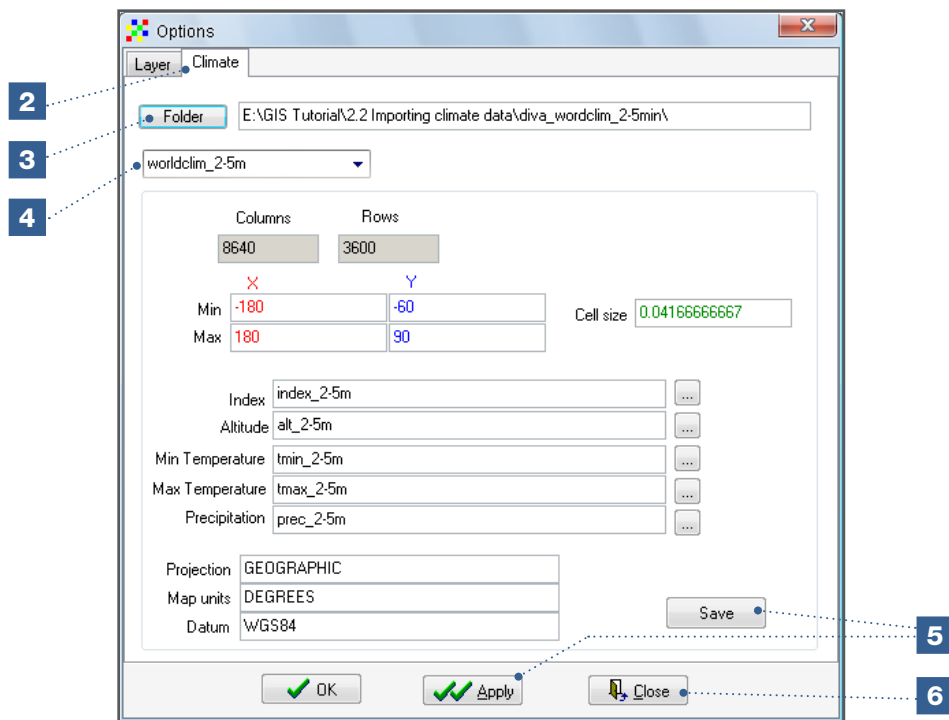
Steps:

1. Copy or extract climate data from the compressed file, *diva_worldclim_2-5min.zip*, to a folder on your computer. In this example, the data are saved using the file-path: *C:\Program Files\DIVA-GIS\environ*.
2. Go to *Tools/ Options* in DIVA-GIS and select the *Climate* tab.



3. Under the *Folder* box, indicate the location of the folder containing the climate data you would like to import. In this example, the data are found in the folder: *C:\Program Files\DIVA-GIS\environ*.
4. Indicate *worldclim_2-5m* as the climate database.
5. Click *Apply* and *Save* to make this the default database for the analysis of climate data in DIVA-GIS.
6. Click *Close* to close the Options window.

7. Return to *Tools/Options/Climate* to view the climate data selected and to check if the climate data have, indeed, been added.

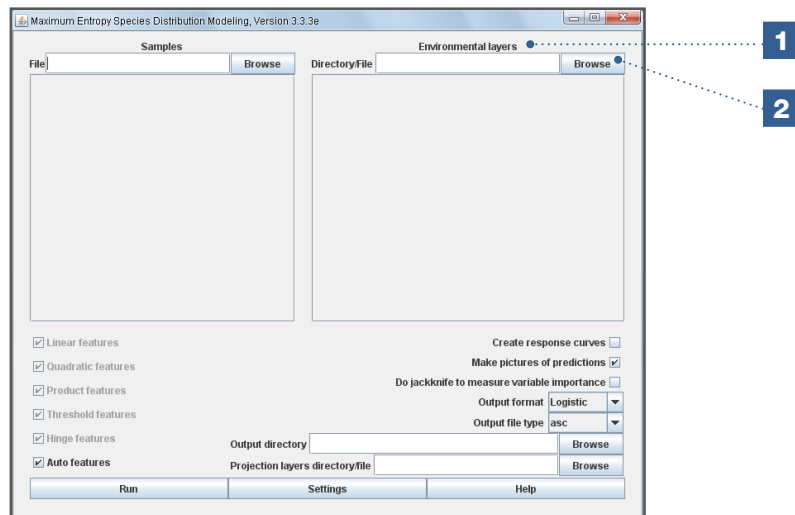


2.2.2. How to import climate data in Maxent

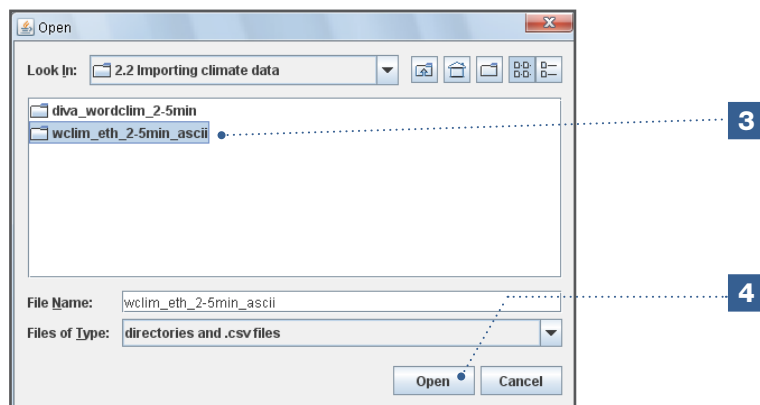
In order to ensure a smooth process when modelling data in Maxent, it is recommended to use environmental raster data in the ASCII format (*.asc). ASCII files (*.asc) can be created in DIVA-GIS under the *Data/Export Gridfiles* option. Analysis 3.1.5 illustrates how to prepare ASCII climate data for a specific study area. Section 6.2 explains how to use Maxent with the 19 bioclimatic variables for the prediction of potential species' distribution. For the example outlined below, the 19 rasters in ASCII format (*.asc), located in the *wclim_eth_2-5min_ascii* folder, are used.

Steps:

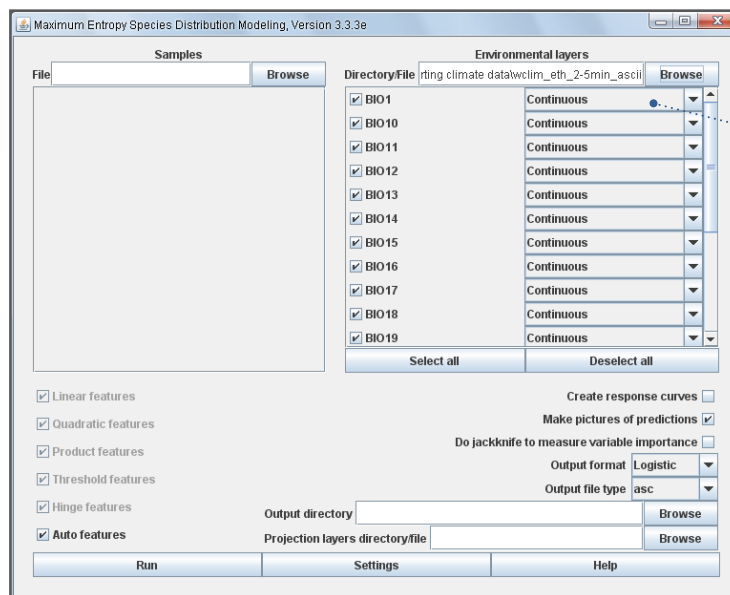
1. Open Maxent and indicate in *Environmental layers* the raster files to be included.
2. Use the *Browse* button to locate the folder: *wclim_eth_2-5min_ascii*.



3. Once you have located the folder, select it by clicking only **once** (do not double-click or open the folder).
4. Click *Open*.



5. All the raster files stored in the folder will be automatically shown in the Maxent interface.



The environmental layers are now ready to be used for species distribution modelling analyses. This is explained in Section 6.2.

2.3. Sources of spatial and other relevant data

There are an increasing number of organizations making spatial data publically available, often within the context of a network. The internet has made it possible to easily share and download data. This section lists key online resources where environmental, geographic and passport data can be retrieved.

Species presence points

- The Global Biodiversity Information Facility (GBIF): <http://www.gbif.org/>
- All separate data providers to the GBIF can be found at the following site: <http://data.gbif.org/datasets/>

Georeferencing and country level data

- Biogeomancer: <http://classic.biogeomancer.org/>
- Geolocate: <http://www.museum.tulane.edu/geolocate/>
- Geonames: <http://www.geonames.org>
- Google Geocoder: <http://code.google.com/apis/maps/index.html>
- Country-level data: <http://www.diva-gis.org/gdata>

Taxonomic information

- Tropicos, Missouri Botanical Garden: <http://www.tropicos.org/>
- GRIN Taxonomy for Plants: <http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl>

Environmental data

- Worldclim - Current, future and past climate data: <http://www.worldclim.org/>
- World soil information (ISRIC) data: <http://www.isric.org/UK/About+Soils/Soil+data/>
- Digital elevation data (STRM 90 meters): <http://srtm.csi.cgiar.org/>
Downscaled GCM Data Portal (to download future climate data): <http://gisweb.ciat.cgiar.org/GCMPPage>

Land cover

- Global Land Cover Facility: <http://glcf.umiacs.umd.edu/index.shtml>
- Global Land Cover 2000: <http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php>
- World Database on Protected Areas (WDPA): <http://www.wdpa.org/>
- World Wildlife Data and Tools: <http://www.worldwildlife.org/science/data/item1872.html>

Miscellaneous

- FAO's Geonetwork: <http://www.fao.org/geonetwork>
- CGIAR Consortium for Spatial Information: <http://csi.cgiar.org/>
- DIVA-GIS: <http://www.diva-gis.org/Data>

References

Busby JR. 1991. BIOCLIM a bioclimatic analysis and prediction system. In: Margules CR, Austin MP, editors. Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Canberra. pp. 64–68.

Hijmans RJ, Cameron SE, Parra JL., Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965–1978.

Chapter 3

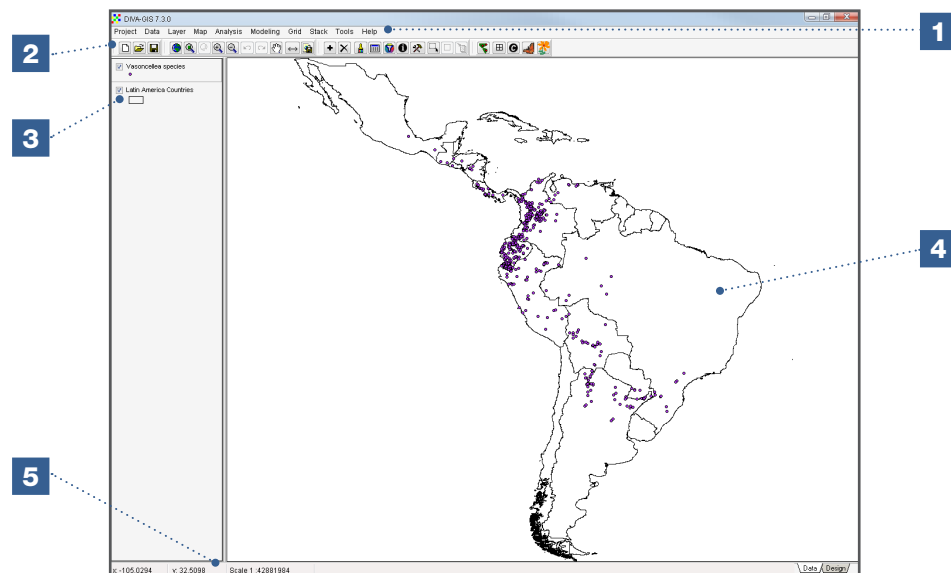
Basic elements of spatial analysis in DIVA-GIS

This chapter illustrates how to use basic tools in the DIVA-GIS programme to carry out common spatial analyses. If you would like to learn more, please consult the DIVA-GIS Operating Manual, available online at: <http://www.diva-gis.org/documentation>.










3.1. Visualization in DIVA-GIS

Before starting the visualization processes, it is important to know the five basic sections presented on the DIVA-GIS work screen.

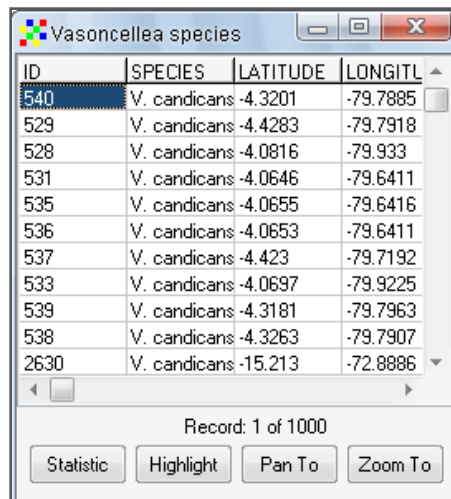
1. **Menu bar:** Facilitates access to all DIVA-GIS commands
2. **Toolbar:** Provides quick access to most commonly used tasks
3. **Legend:** Lists all layers of the current map. The selected map is highlighted. Layers can be visualized or hidden by checking or un-checking the boxes in front of each layer name
4. **Map:** Visualization of the current map
5. **Status bar:** Indicates the coordinates where the cursor is located, the map's scale and the raster's position and value.



The following table shows key navigation commands that allow you to visualize and explore various aspects of the maps. Each command is activated by clicking on the respective button. Depending on the information attributed to each map, certain buttons may be inactive.

Button	Commands
	<i>Zoom in</i> (zooms in on a point by left-clicking the mouse or zooms in on an area by drawing a rectangle while holding down the left mouse button)
	<i>Zoom out</i> (zooms out from a point by a left-clicking the mouse)
	<i>Pan</i> (moves the visible zone of the map by holding down the left mouse button and moving the mouse)
	<i>Zoom to active layer</i> (zooms out to the extent of the currently active layer)
	<i>Zoom to full extent</i> (zooms out to the maximum extent of all layers)
	<i>Information</i> (provides information on the element identified by the mouse in the layer selected)
	<i>Remove Layer</i> (eliminates any layer that has been selected)
	<i>Table button</i> (see further explanation below)
	<i>Distance button</i> (allows you to calculate distance between different features)

The *Table* button allows you to access the table of attributes of a vector file (these attributes can be explored with the *Statistic*, *Highlight*, *Pan to* and *Zoom to* buttons.) The table is a 'read only' file and cannot be changed. You must access the original file used to generate the vector file (see 2.1.2) in order to modify the data. It is recommended to prepare a new spreadsheet file with the modified data and not to alter the original file. In this way, if a modification is invalid, you can still return to the original data.



ID	SPECIES	LATITUDE	LONGITL
540	V. candicans	-4.3201	-79.7885
529	V. candicans	-4.4283	-79.7918
528	V. candicans	-4.0816	-79.933
531	V. candicans	-4.0646	-79.6411
535	V. candicans	-4.0655	-79.6416
536	V. candicans	-4.0653	-79.6411
537	V. candicans	-4.423	-79.7192
533	V. candicans	-4.0697	-79.9225
539	V. candicans	-4.3181	-79.7963
538	V. candicans	-4.3263	-79.7907
2630	V. candicans	-15.213	-72.8886

Record: 1 of 1000

Statistic Highlight Pan To Zoom To

- **Statistic:** in the case of numeric values, the *Statistic* button provides a summary, in an additional window, of the basic statistics of numerical variables.
- **Highlight:** highlights the presence point/line/polygon selected for a few seconds.
- **Pan to:** the presence point/line/polygon you wish to view may not be in the section


of the map initially displayed. This tool moves the map (with the same zoom) to the place where the point is found.

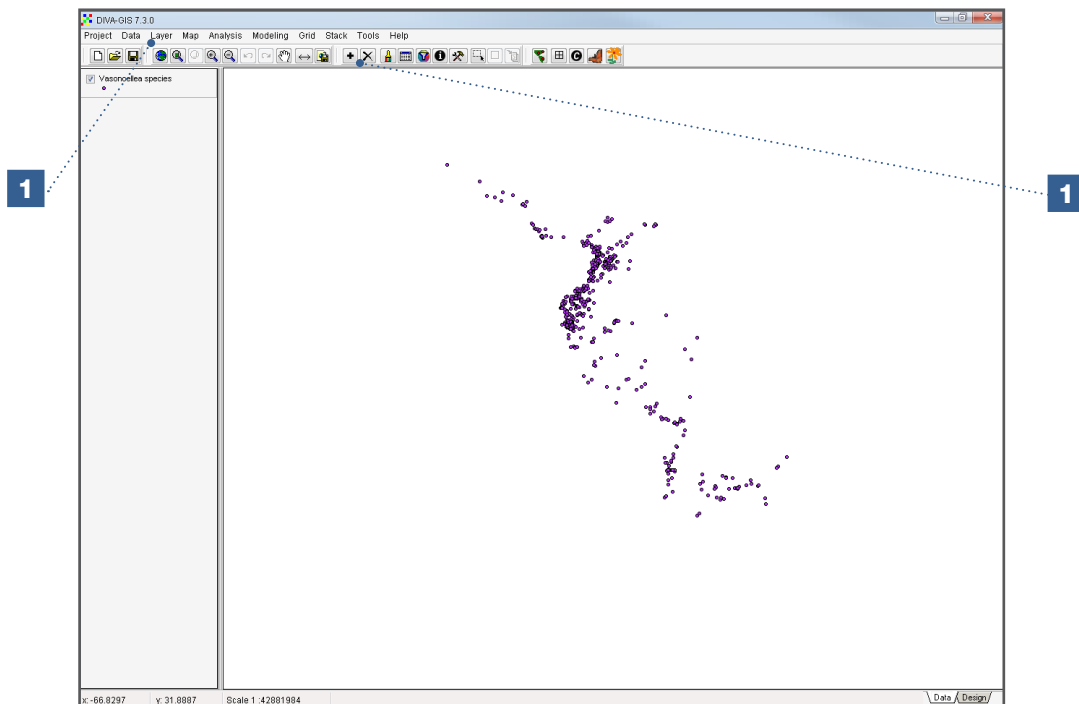
- **Zoom to:** this option also moves the presence point/line/polygon to the centre of the displayed map, but increases the zoom.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel 	<p>Data Files:</p> <p>Folder 3.1 Basic elements</p> <ul style="list-style-type: none"> • <i>Vasconcellea species (shp, shx, dbf)</i> • <i>Latin America countries (shp, shx, dbf)</i> • <i>COL_ADM1 (shp, shx, dbf)</i> • <i>Mean temperature Latin America 10 min (grd,gri)</i> • <i>Precipitation Latin America 10min (grd,gri)</i> • <i>World_adm0 (shp, shx, dbf)</i> <p>For the following analyses, you need to have the 2.5 min worldclim climate data imported in DIVA (cf. Section 2.2)</p>

3.1.1. How to perform basic visualizations using vector files

Steps:

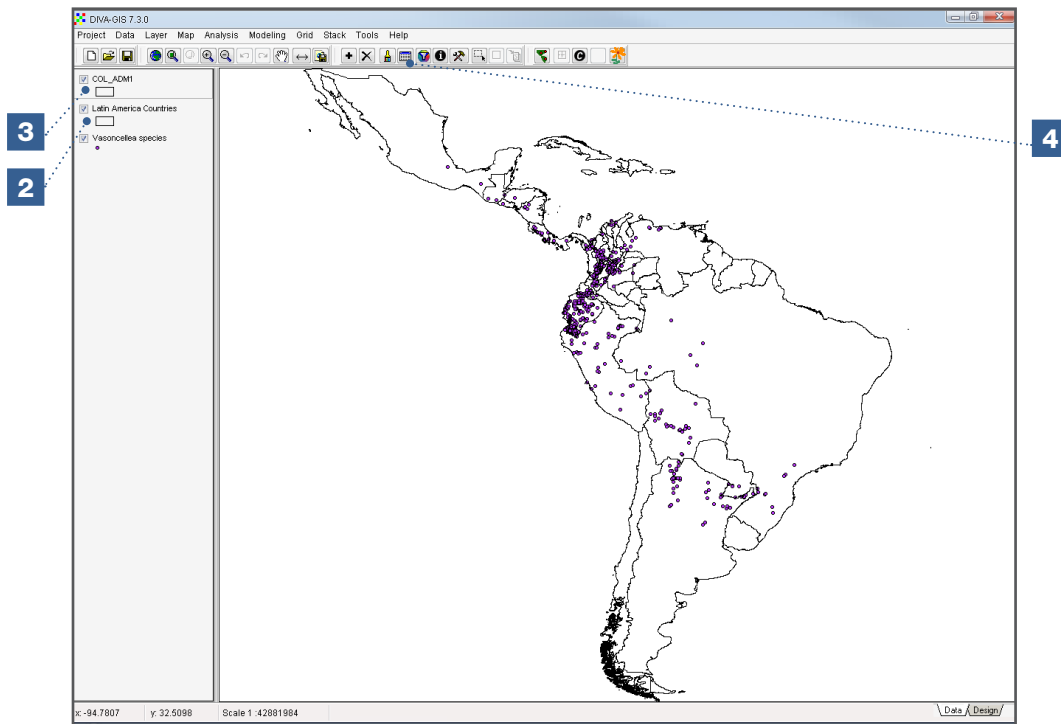
1. Open the *Vasconcellea* (*Vasconcellea species.shp*) layer of points by using the  icon or the menu option: Layer/Add Layer. This simple display of presence points does not provide much information; it is therefore necessary to add more layers to complement these data points.



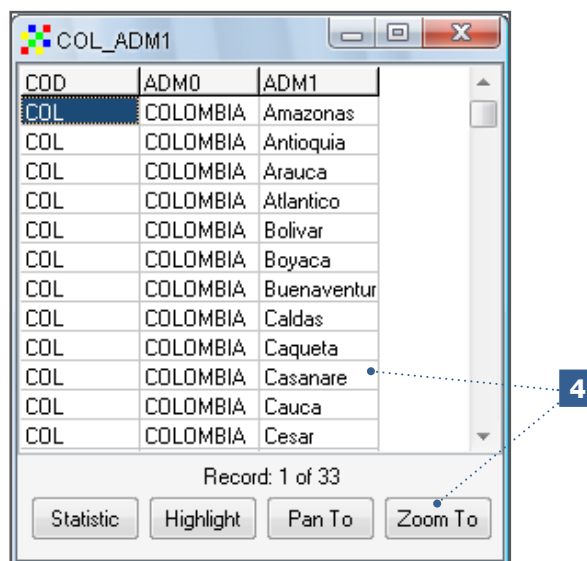
2. Add the layer which includes the countries of Latin America and the Caribbean (file: *Latin America Countries.shp*). The layer should display immediately. Layer names are

shown in the left column as they are added. Check (or uncheck) those boxes next to the layers you are interested in visualizing (or hiding).

- Colombia is used here as an example to provide further detail during this analysis. Add the layer with Colombia's departments (*COL_ADM1.shp*).



- Information on Colombia's administrative divisions can be obtained by selecting the newly added layer (click on it) and clicking on the table icon. This opens the table information. We can now find specific features in the *.shp file. For example, find out where the Casanare Department is located using the buttons *Highlight*, *Pan To* and *Zoom To*.

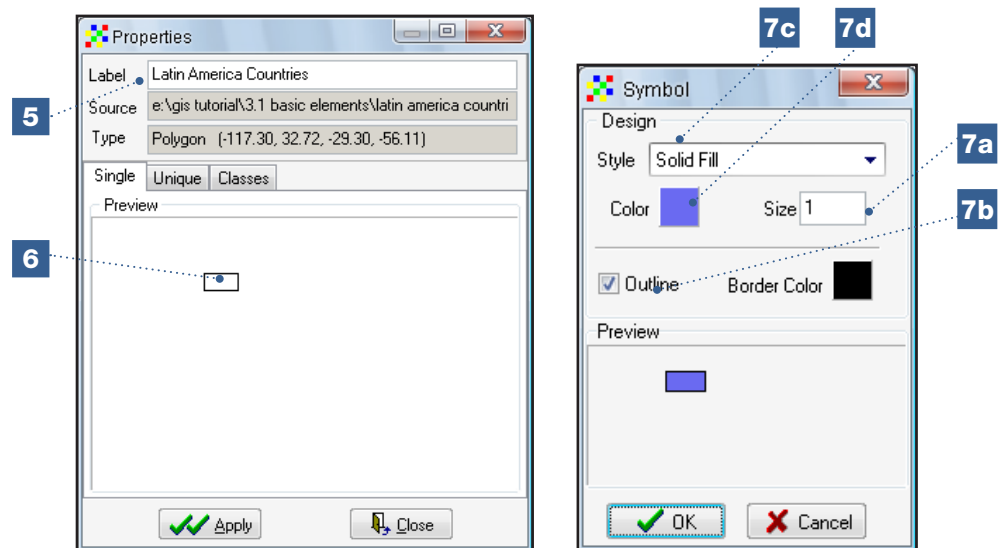


- Now hide (uncheck) the layers of *Vasconcellea* species and remove the layer of the Colombian departments. Expand the map so it shows all *Latin American countries*.

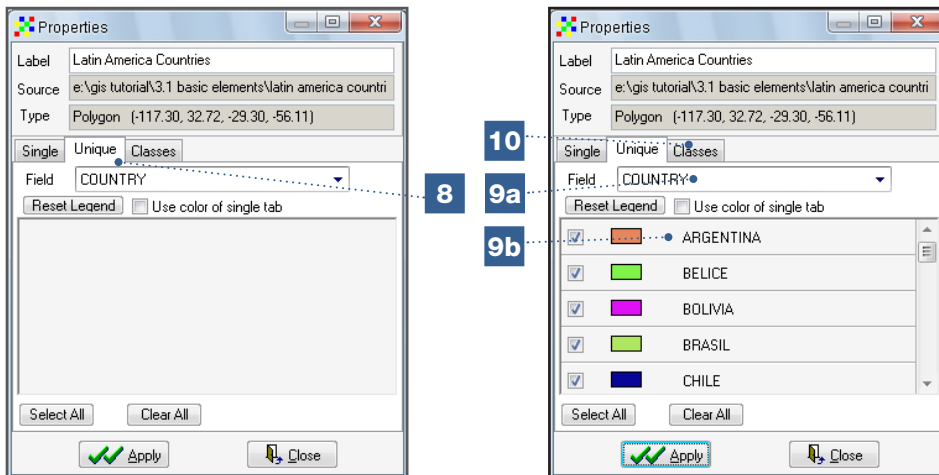
To change a layer's legend, double-click on the layer to be modified. For this analysis, click on the layer: Latin America Countries. The *Properties* window will immediately be displayed, showing the *Single* tab as the first option. This window will allow you to make changes to the legend.

6. The *Single* tab displays a transparent rectangle. A double-click on the rectangle opens the *Symbol* window.
7. The attributes of the *Symbol* option allow you to change certain properties of the polygons:
 - a. Outline thickness
 - b. Outline colour
 - c. Polygon filling style
 - d. Polygon colour.

Try to change the default values of these properties and observe what happens after clicking *OK* in the *Symbol* window and *Apply* in the *Properties* window. For the best visualization, the recommended filling style is: *Solid Fill*.



8. It is best to use different colours for different attribute values. In the *Properties* window, select the *Unique* tab.
9. Under the *Field* tab, select the field to be used to create the layer's new legend. It is important to select a field with a limited number of classes. If you use a field containing numeric values (e.g. area), you may cause DIVA-GIS to hang or lag (since the programme will take each value as a different class). For this exercise, select the field: *Country*, [9a], and click on the *Reset Legend* button. Finally, click on *Apply*. The legend for the selected layer will be displayed. (Colours can also be changed in this menu by double-clicking on the rectangle indicating the colour for each class [9b].) In order to assign different colours to each class, it is important to have a fill style selected under the *Single* tab (e.g. *Solid Fill*).



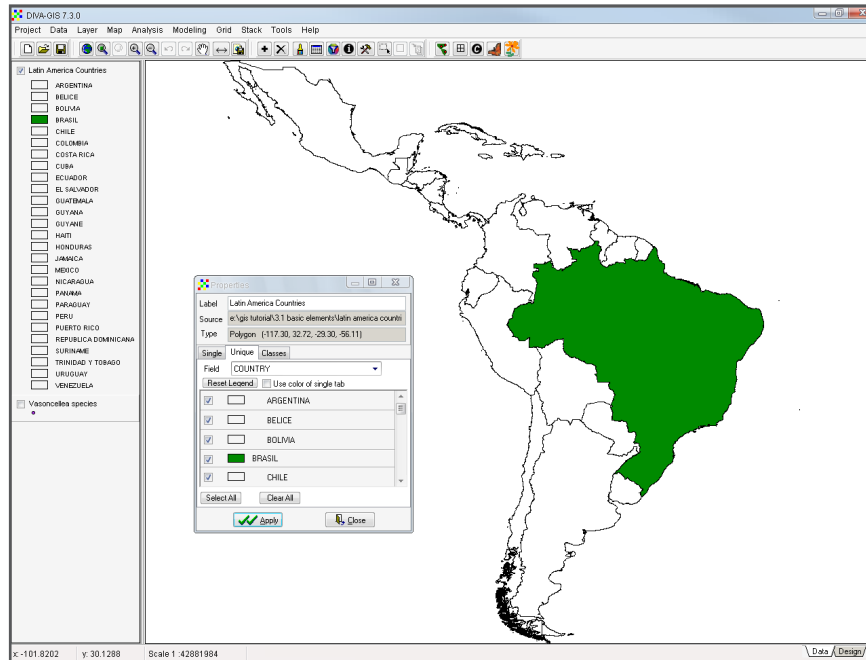
10. A similar process is used for numeric values under the *Classes* tab. (This is not a commonly used process and is therefore not included in this manual.)

If you have followed the steps correctly, the map below should be displayed (colours may be different):

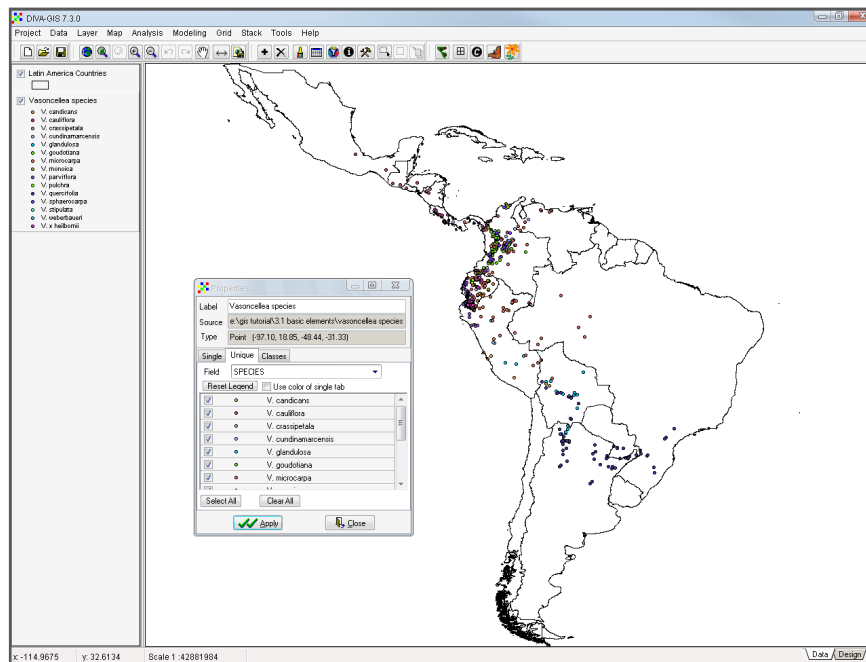


Individual Task: Make Brazil green and eliminate the colour in the remaining countries.

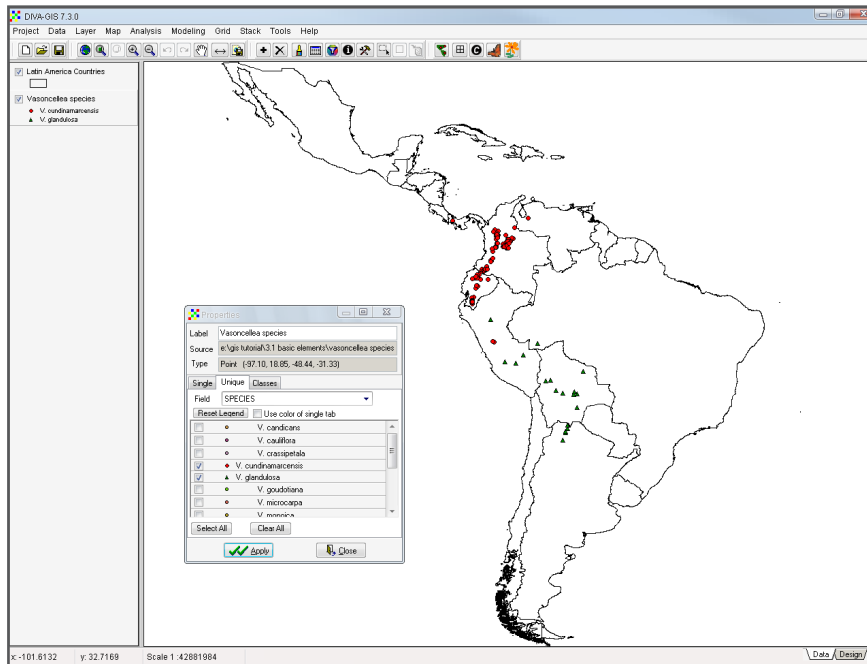
Result:





11. Now, remove the countries legend (return to the *Single* tab and eliminate *Solid Fill*) and visualize the layer with the *Vasconcellea* points again. This time, assign a different colour to each species following the procedure explained in the previous steps.

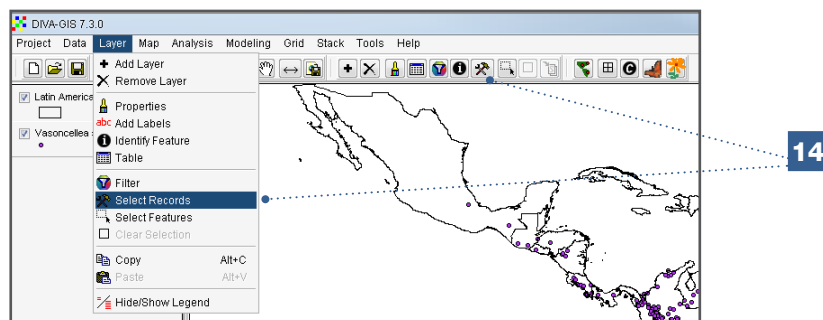


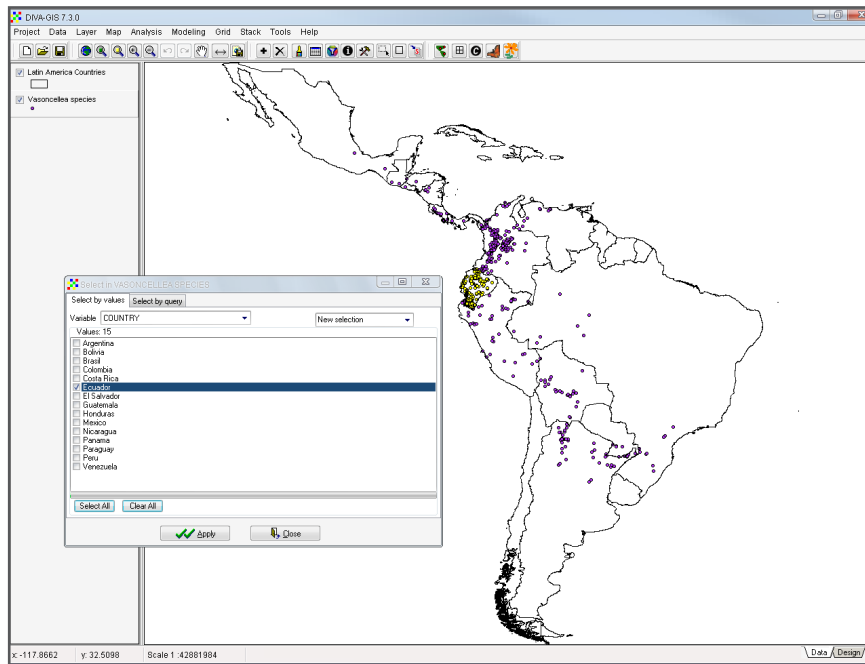
12. Using the options in the *Properties* window, assign a red circle to observations for the species *V. cundinamarzensis* and a green triangle for the species *V. glandulosa*. The other species should remain invisible.



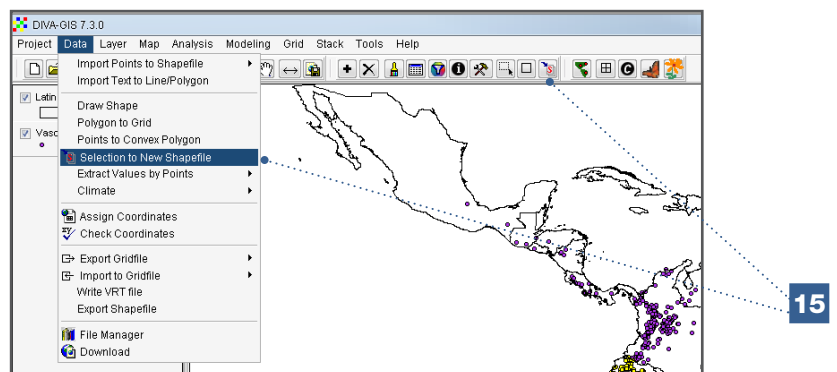
It is important to realize that a selection at the legend level will not change the content of the vector files. To carry out an analysis of a subset of data, you must first select the subset and save it as a different vector file. Return to the *Single* tab in the legend and select a single symbol for all points. All observations are again visualized with a single symbol.


13. Now, select (not display) a specific group of *Vasconcellea* species using either of the two attribute selection alternatives provided by DIVA-GIS in the *Layer* menu: *Select Records* or *Select Features* (these options are also available for quick access in the toolbar: buttons  and , respectively).
14. Click on *Select Records* to select groups according to a specific variable (*Select by values*) or by a combination of variables (*Select by query*). For this analysis, all occurrences of *Vasconcellea* in Ecuador have been selected. Yellow points indicate the group selected.

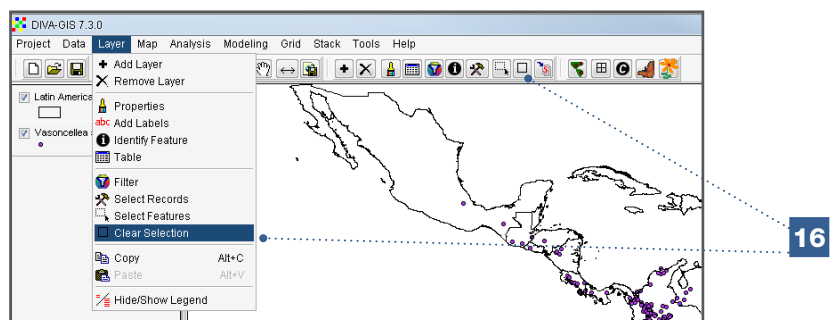




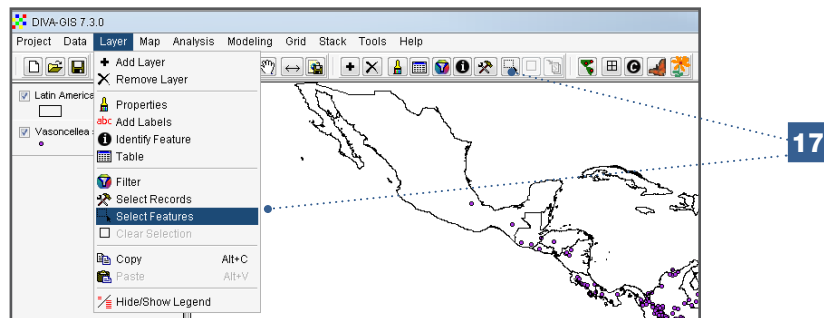
- The next step is to save the selection as a new vector file. In the *Data* menu, click on *Selection to New Shapefile* to save the selection (*Vasconcellea* species in Ecuador) as an additional layer, which can now be seen on the legend bar. Hide this new layer.



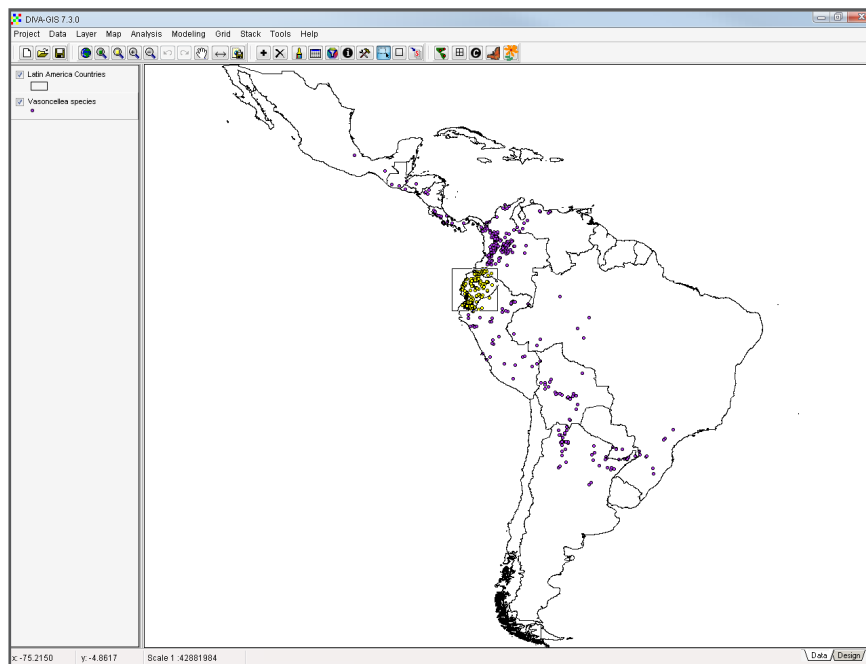
- Sometimes, it is important to remove selected features. To do this, select the layer which holds the selected features and go to *Clear Selection* in the *Layer* menu, or click on the corresponding button in the toolbar (). The selected points (in yellow) are no longer highlighted.



17. Now, make a new selection by selecting the points directly on the map, using the *Select Features* option in the *Layer* menu. For this analysis, continue to work with the *Vasconcellea* species layer (keep it selected).



The cursor will immediately take the shape of a cross (+), allowing you to manually select a group of points (keep holding the left mouse button down and drag the cursor over the group of points you would like to select). The selected points will change colour. Once again, the selection can be saved as a new layer (Step 15).



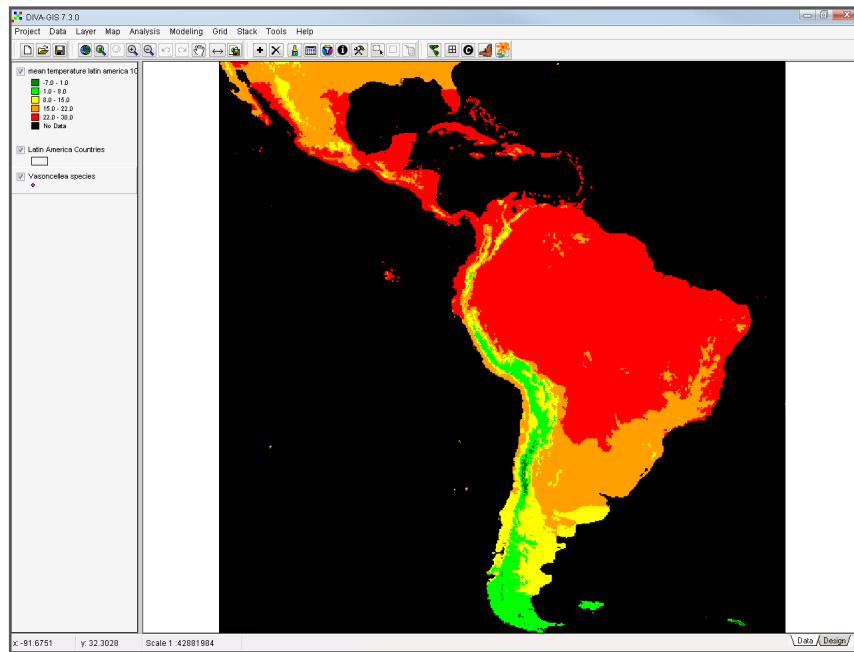
3.1.2. How to perform basic visualizations using rasters

As mentioned in Chapter 2, DIVA-GIS also manages raster data (raster files have the extension *.grd in DIVA-GIS). Actually, most of the analyses outlined in this manual will result in rasters.

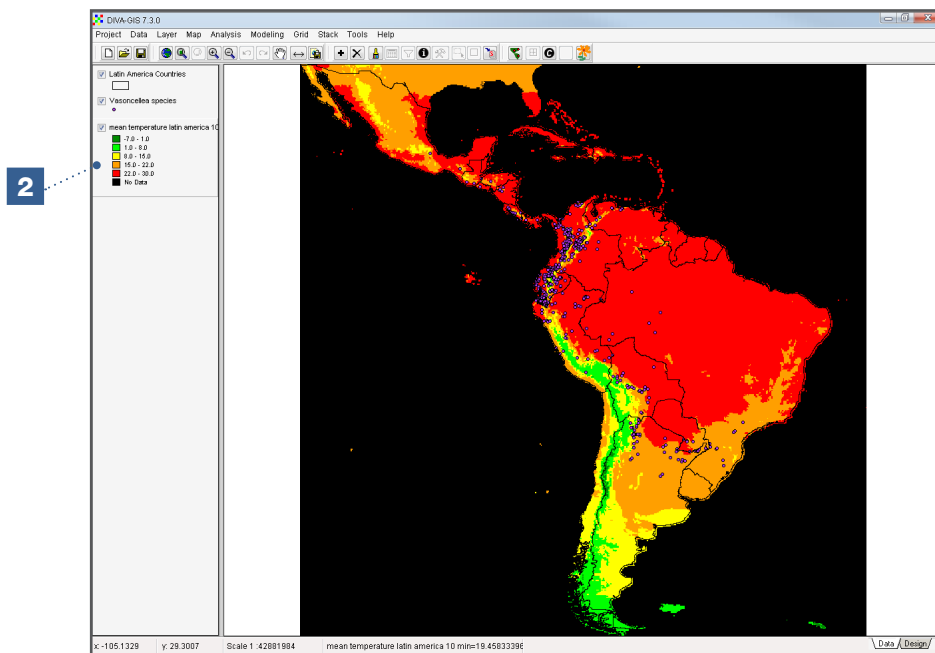
Steps:


1. Open the *Vasconcellea* species and the *Latin America Countries* layers. Add the layer with the values for annual maximum temperatures in Latin America (file: *Mean*

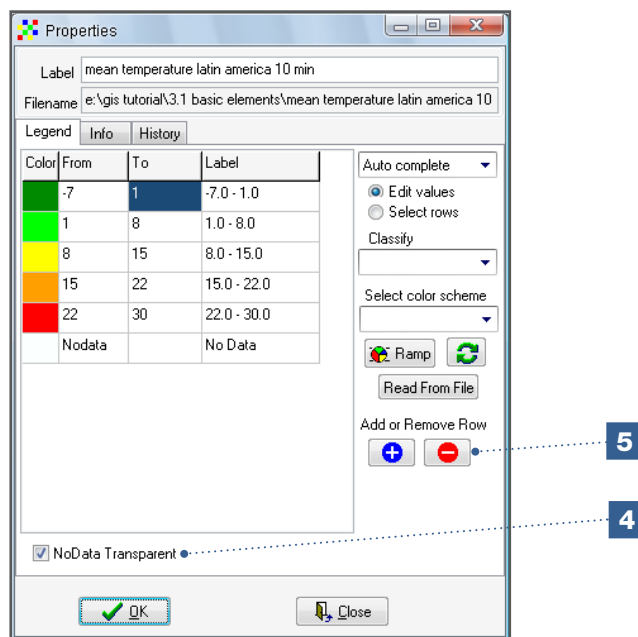
temperature Latin America 10 min.grd). This layer corresponds to a raster, which is added to the legend panel in the same way as previously shown with a vector file.



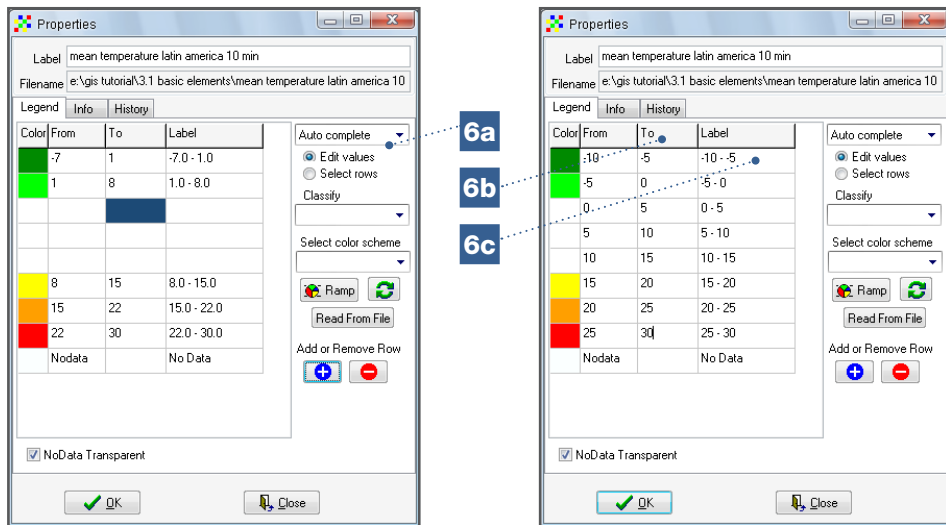
2. Notice the new layer covers the previous ones (i.e. country borders and observation points). The order in which the layers are listed in the legend column corresponds to the order in which they are displayed (superimposed). To modify the order, reorganize the layers by dragging them into the desired position. To continue, relocate the temperature layer in the bottom position in order to allow the other layers (countries and points for species occurrence) to be visualized on top.



3. The different classes of temperature are expressed as ranges (this information is displayed in the legend). As a default, DIVA-GIS uses five classes with equal ranges, which are often not the most suitable means to visualize results. Through the *Properties* menu, the classes can be adjusted. This menu is accessed by double-clicking on the layer.
4. Select the *NoData Transparent* box at the bottom of the window, which will remove the black colour for those cells with no data.
5. Proceed to assign new classes (at five degree intervals), starting at the minimum temperature of -10 °C until you reach the maximum temperature of +30 °C. This classification will result in eight classes (a range of 40 degrees divided by 5). Therefore, you will need to add three more rows in the default table. The plus and minus buttons () allow you to add or delete classes, as required.



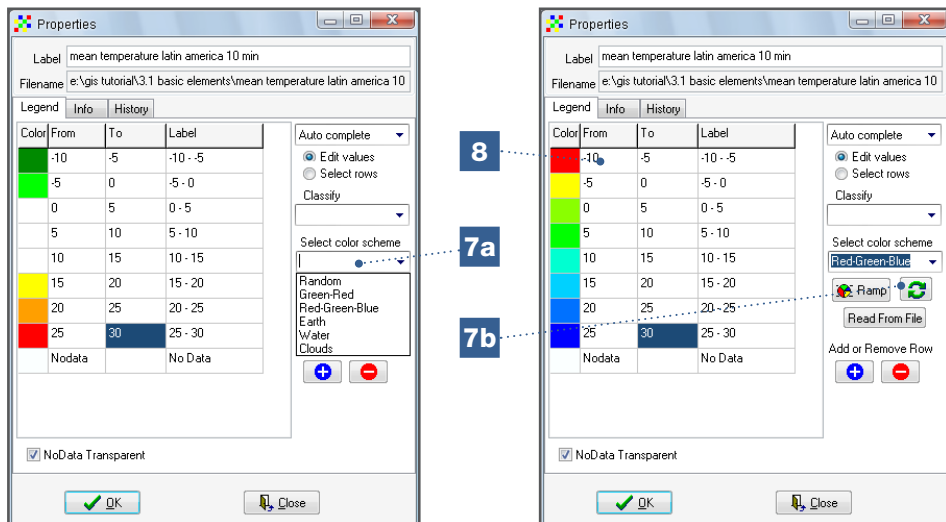
6. Check the following parameters before reorganizing the ranges:
 - a. Select the commands *Auto complete* and *Edit Values* (shown in the left column) to allow the entering of values in the *From* and *To* columns.
 - b. The values of the new classes must be entered in the boxes, starting with the upper section of the table (*From*: -10; *To*: -5).
 - c. After entering the first value, the cells (as well as their labels) in the *From* column should change automatically for the next row; you only need to enter values in the *To* column.



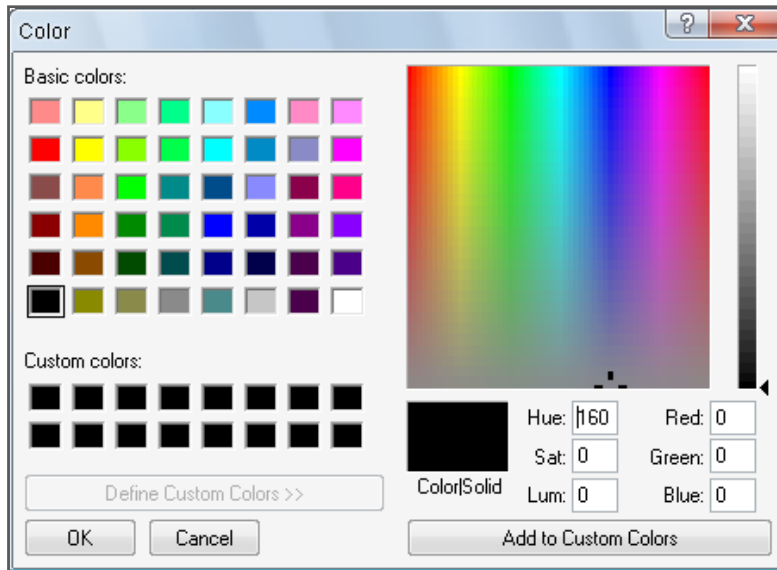
Note

It is also possible to change the values in the *From* and *To* columns manually. To do this, select the *Manual* option in the first box (cf. 6a). This will allow you to change the values in these columns. By choosing this option, there is the risk that some values might be forgotten and will therefore not be displayed on the map.

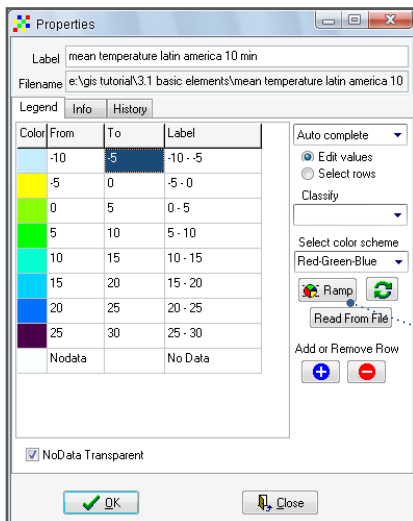
- The colours of each temperature range can be changed by using the *Select Color Scheme* command and selecting the options available, such as *Red-Green-Blue*. By doing this, the colder zones will display in red and the hotter in blue (7a). The order of the colours can be inverted by clicking on the button with two arrows, highlighted below (7b).



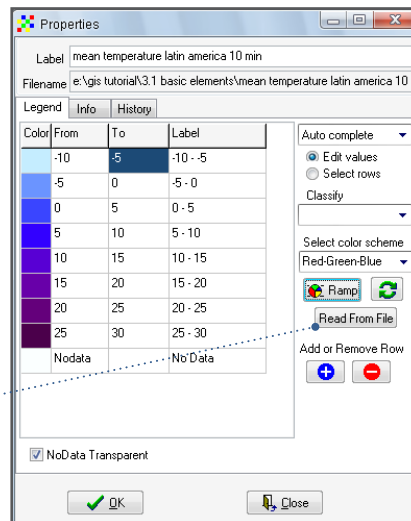
- You can also personalize the colours, if preferred. Double-click on the first colour and choose the desired tone (you can select either *Basic colors* or *Custom colors*). Repeat the process for the colours in each class.



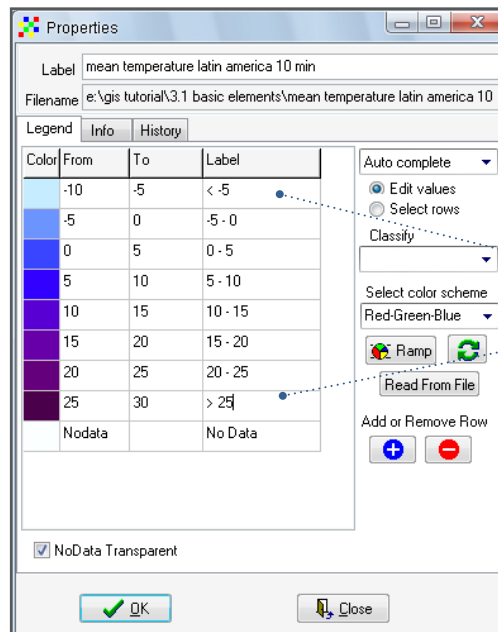
- The *Ramp* button allows you to select only extreme tones; this tool will automatically suggest the intermediate scale of tones. For this analysis, select the extreme tones: pale blue and burgundy red.
- Some circumstances require legends to be consistent among different maps. To use the legend information from an existing raster, click the *Read From File* tab, and indicate file path.



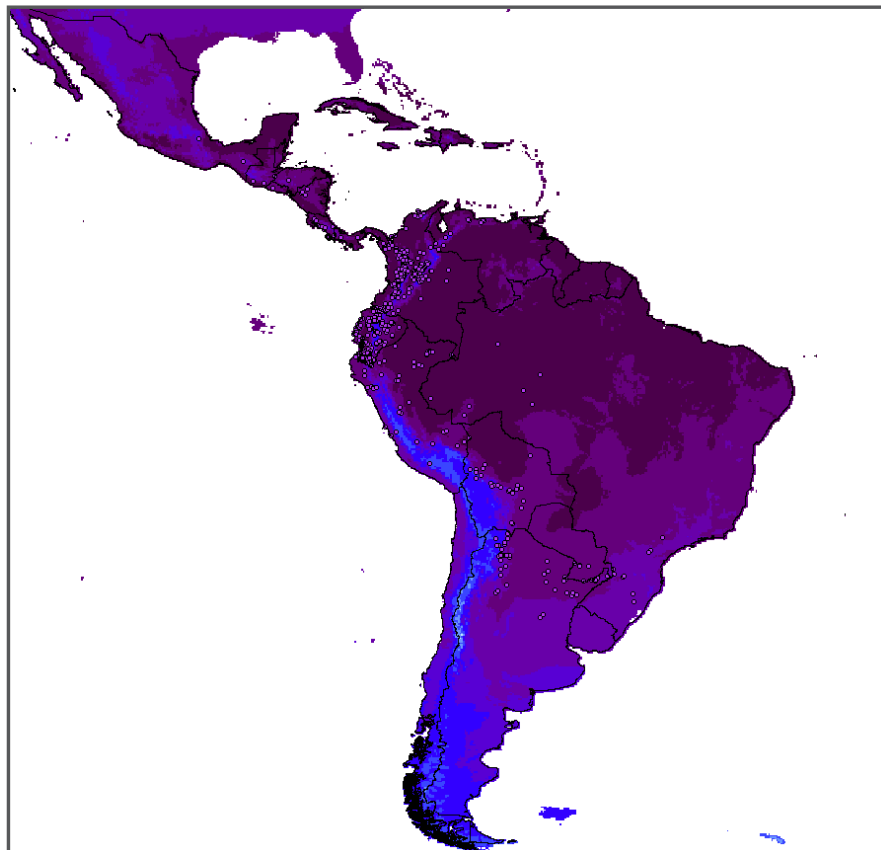
9
10



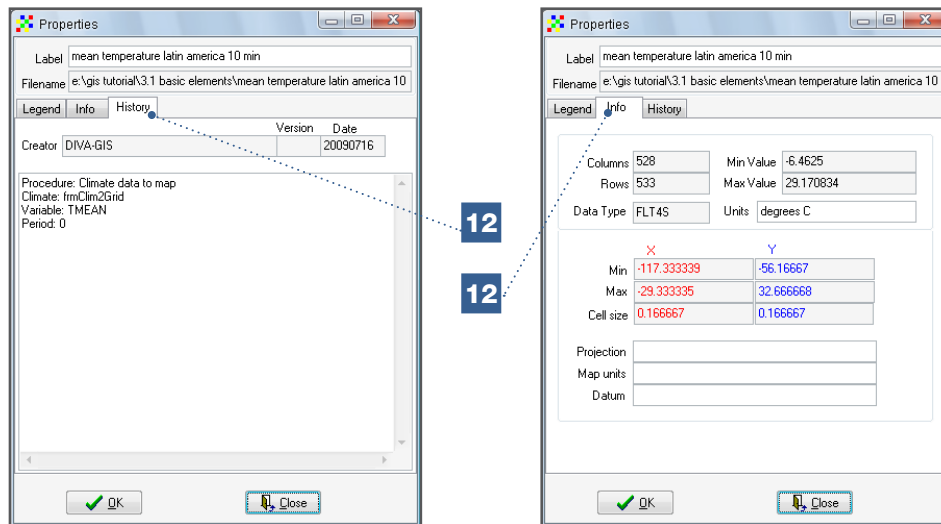
11. The label (text) describing each class can also be manually modified. For this analysis, change the label for the first class: *From* [-10 – -5] *To* [< -5] and for the last class *From* [25 – 20] *To* [> 25]. Click *OK* to illustrate these changes on the map.



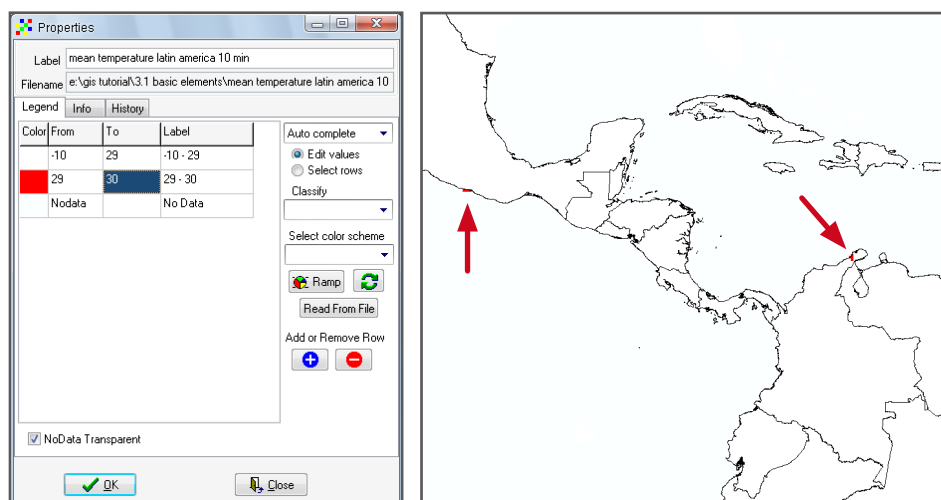
If you have followed the procedure correctly, the map displayed should be similar to the one below:



12. The *Properties* button also includes additional tabs for further information. Under the *History* tab you will find the information on how the layer was generated (which can be useful in the case of an unexpected outcome or in order to find details on how a previous analysis was carried out), while the *Info* tab offers relevant information on the raster, including resolution and maximum/minimum values.



13. In this analysis, clicking on the *Info* tab allows you to see that the maximum temperature is 29.17 °C, but the exact place on the map corresponding to this extreme value is difficult to locate.
14. Try to find the sites with the highest temperature (29.17 °C) in Latin America. To do this, click on the *Legend* tab and create a unique class that contains temperature data superior to 29 °C. Select *red* as the colour for this range. The resulting map should show a few cells in Mexico and Colombia. Now you must check each cell to identify which has the highest value. The status bar at the bottom of the screen gives the exact value for each cell. The hottest place in Latin America is located in southern Mexico.

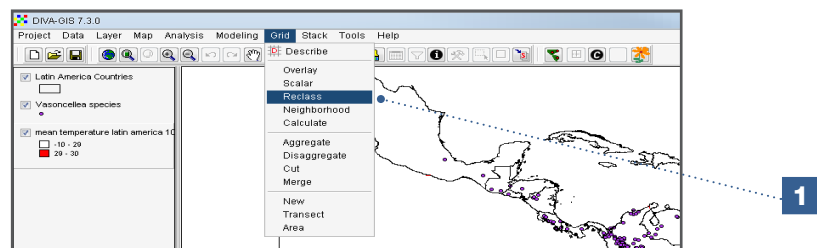


3.1.3. How to combine rasters

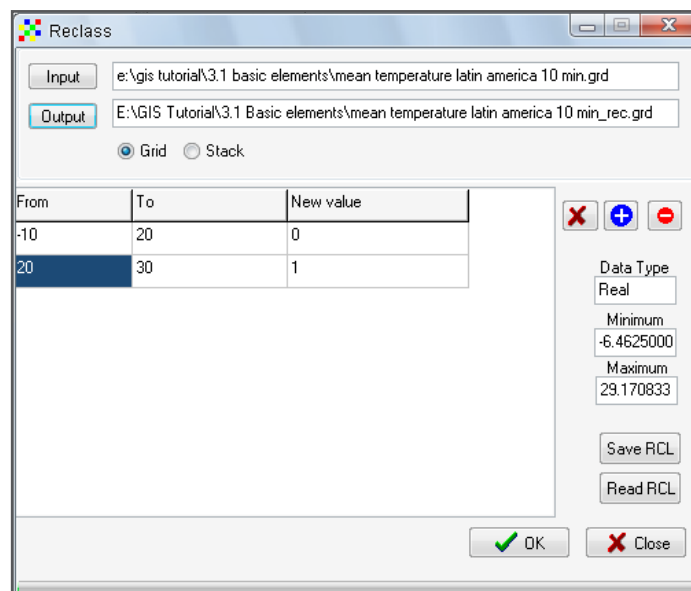
Similar to making changes in the legends for vector layers, changing the legend in rasters only affects the display, not the file's original information. Sometimes it is relevant to combine selected zones of different rasters; for example, the hot areas from a temperature raster with the dry areas from a precipitation raster. The original rasters contain information on both the selected and non-selected zones, which makes them of little use for such combinations. Therefore, new rasters of the selected zones need to be created to combine the information of interest.

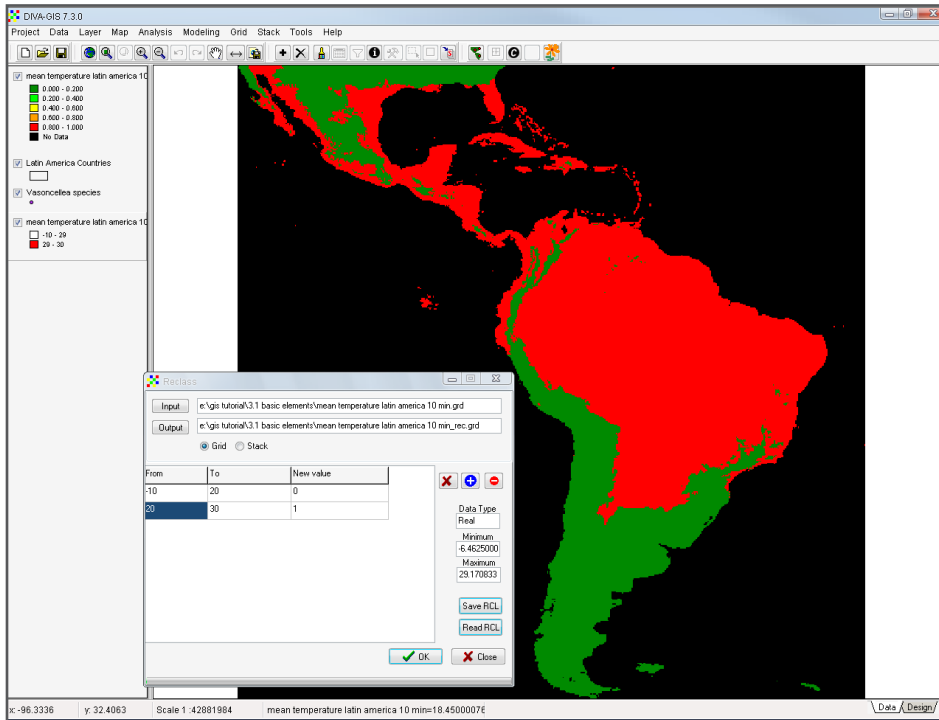
Steps:

1. For this analysis, select all areas in Latin America with an average temperature of > 20 °C and precipitation < 1000 mm. Start with the temperature (layer: *Mean temperature Latin America 10 min.grd*), selecting all areas with averages greater than 20 °C. In the menu, go to *Grid/Reclass*.

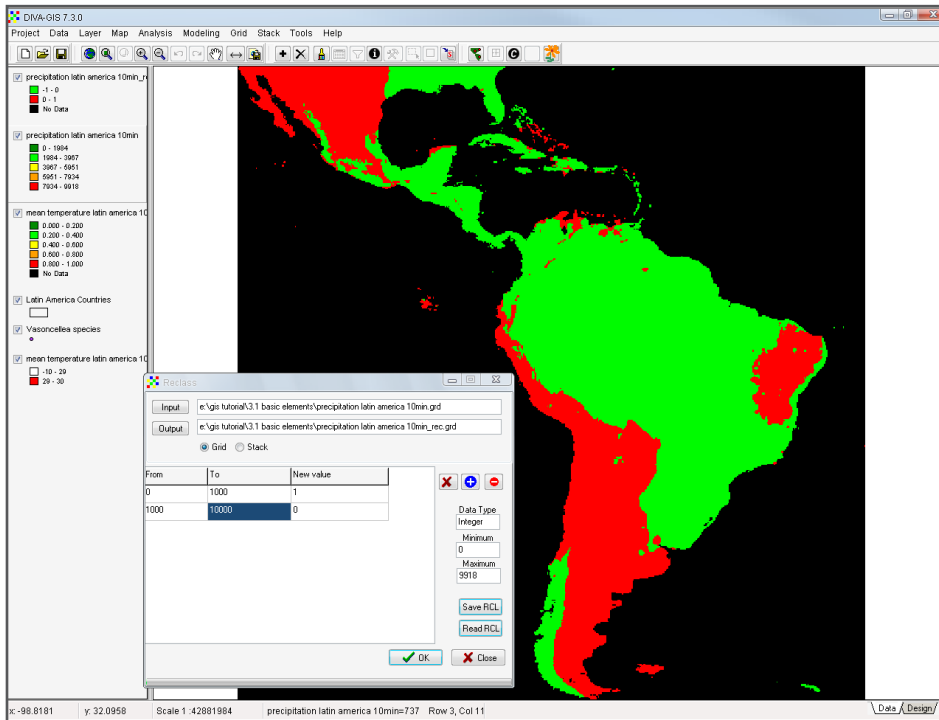


2. Change the values into classes (make sure to cover the whole range from minimum to maximum value), assigning a value of zero (0) to those cells with values less than 20 °C and a value of one (1) to those equal to or greater than 20 °C. Assign a name to the new raster generated using the *Output* button. The resulting display should only include cells selected with average temperatures equal to or greater than 20 °C.

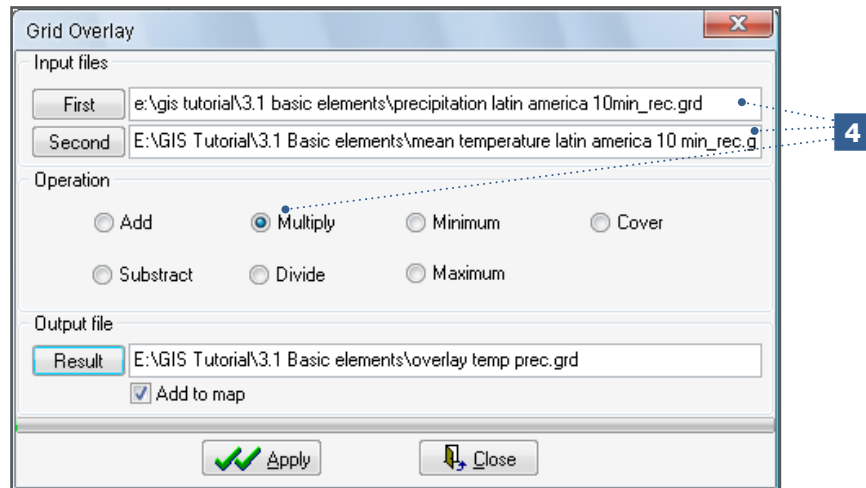




3. Add the precipitation raster (*Precipitation Latin America 10min.grd*). By using the *Reclass* option, you can select the cells where precipitation is below 1000 mm (see Step 1).



4. DIVA-GIS allows you to combine rasters using the *Grid Overlay* tool. We now need to identify zones with average temperatures greater than or equal to 20 °C and annual precipitation below 1000 mm. Under *Input* files, select the two layers generated in the previous steps and then click *Multiply*. Please see the following table to understand how the calculation works.



	Average temperature ≥ 20 °C and Precipitation < 1000 mm	Average temperature ≥ 20 °C and Precipitation > 1000 mm	Average temperature < 20 °C and Precipitation < 1000 mm	Average temperature < 20 °C and Precipitation > 1000 mm
Temperature raster cell value	1	1	0	0
Precipitation raster cell value	1	0	1	0
Combination (multiplication)	1	0	0	0

The result is a raster where the cells, with a combined value of one, indicate the areas that comply with the two conditions.



There are many other options to visualize and manipulate vector and raster layers. Some important options include:

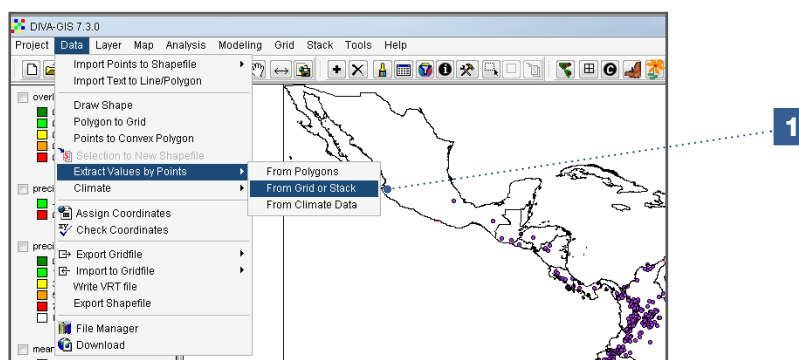
- Selecting *Grid/Aggregate* to combine cells and reduce the resolution of a raster (see Section 6.4) and *Grid/Disaggregate* to split cells to increase the resolution (note that this can give a false sense of precision). The objective of these processes is to ensure two rasters have the same resolution and extent, which is critical when combining them.
- Using *Stack/Calculate* to combine three or more rasters with the same characteristics (in resolution, origin and size) using more complex manipulations (e.g. calculations) than those provided by *Overlay*.
- Clipping part of a raster using *Grid/Cut*. This can be used to create a dataset which includes only the area (known as extent) you are working on. This will result in a reduction of processing time when running an analysis, such as species distribution modelling, which is explained further in Chapter 6. This is particularly important if the dataset you are using is very large.
- Changing the projection of vector files using DIVA-GIS. By selecting *Tools/Projection*, conversions between different projections, such as latitude/longitude and UTM, can be made.

3.1.4. How to extract values from rasters based on presence points data

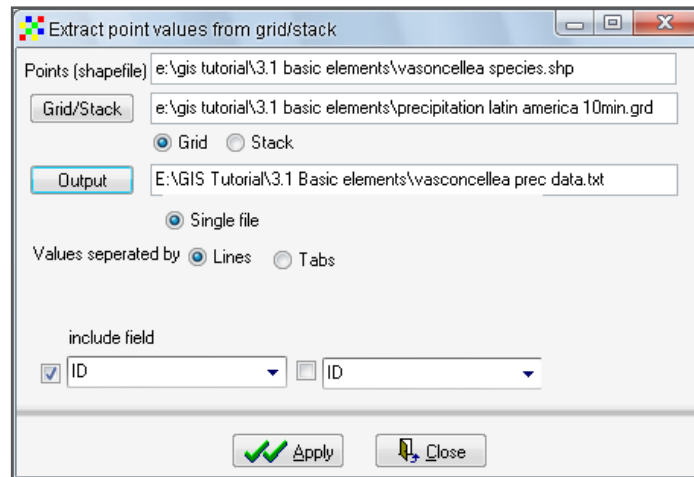
The previous steps detail how to display and manipulate vector and raster layers. It will often be useful to combine vector and raster data based on a common geographical location. In the exercise outlined below, the precipitation value is extracted from the climate raster for each point in the *Vasconcellea* database (vector file) based on its geographical location. Hide all previously generated grid layers.

Steps:

1. Select the layer with the *Vasconcellea* points (*Vasconcellea species.shp*) and go to the *Data* menu. Select *Extract Value by Points* and then select *From Grid or Stack*. A stack is a group of rasters with the same characteristics (in resolution, origin and size).



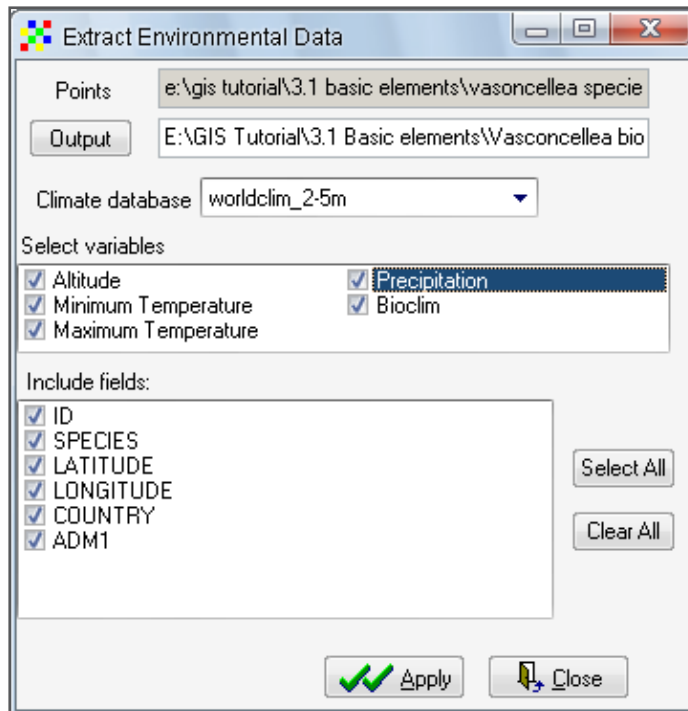
- Select the *Grid* option and mark the raster from which you want to extract the data (this exercise uses the raster file with the precipitation data *Precipitation Latin America 10min.grd*). Leave the default values (*Single File, Lines*) and insert a name for the *Output File* (which will be a text [* .txt] file). You need to use the vector file's identification (*ID*) fields since these are key for combining the generated text file with the data in the observation database (*.xls or *.dbf file).



- When you open the resulting text file in Excel, you will notice that the IDs are combined with the extracted raster values. It is now possible to use the information from both sources, e.g. for ecological niche analyses. This is further explained in Section 6.1.

grid	ID	values
precipitation latin america 10min	540	1139
precipitation latin america 10min	529	1019
precipitation latin america 10min	528	1352
precipitation latin america 10min	531	998
precipitation latin america 10min	535	998
precipitation latin america 10min	536	998
precipitation latin america 10min	537	1019
precipitation latin america 10min	533	1352
precipitation latin america 10min	539	1139
precipitation latin america 10min	538	1139
precipitation latin america 10min	2630	526
precipitation latin america 10min	542	484
precipitation latin america 10min	543	21
precipitation latin america 10min	544	22
precipitation latin america 10min	2628	503
precipitation latin america 10min	2818	2436
precipitation latin america 10min	2913	2912
precipitation latin america 10min	2915	1718
precipitation latin america 10min	2820	3117

4. Finally, if you would like to extract all the climate data, you can repeat this process, but instead of selecting the *Extract Values by Points/From Grid or Stack* option, you will need to select the *Extract Values by Points/From Climate data* option. This can only be done if CLM files (*.clm) with climate data are connected to DIVA-GIS (see steps outlined in Analysis 2.2.1). The resulting file will be a text file which includes all selected climate data, as well as the original data (if selected).



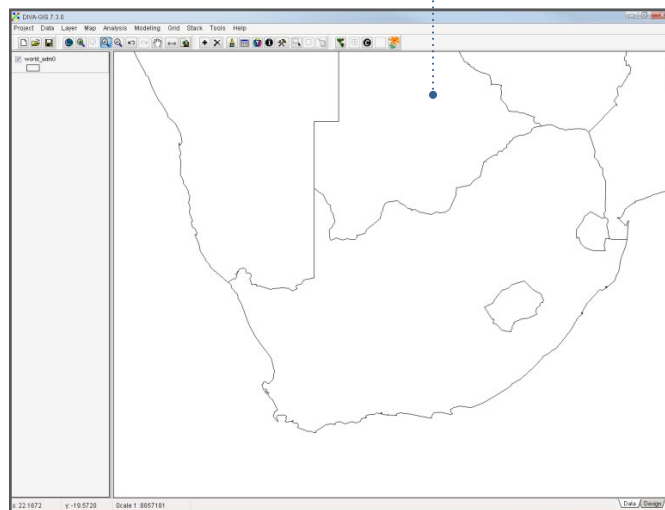
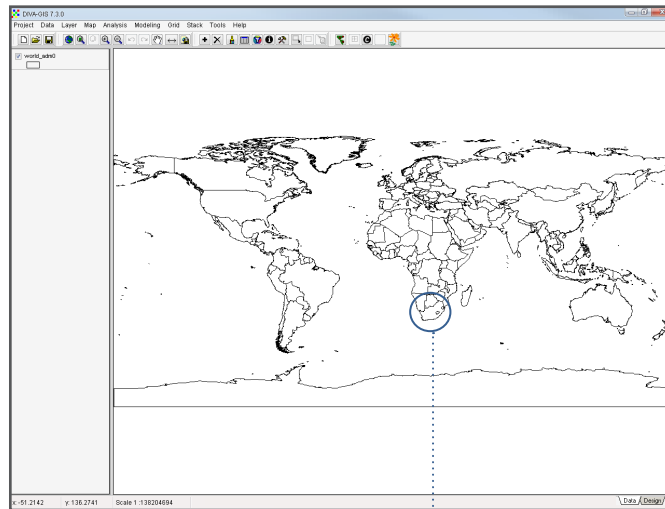
3.1.5. How to create custom-made climate layers

Besides extracting site-specific climate data from the information included in the CLM files (*.clm), DIVA-GIS also allows one to generate parameter-specific climate layers (like those used in Analyses 3.1.2, 3.1.3 and 3.1.4). These rasters are not only useful to gain a better understanding of the climatic conditions in the study area, but can also be used in species distribution modelling (given that all rasters have the same properties).

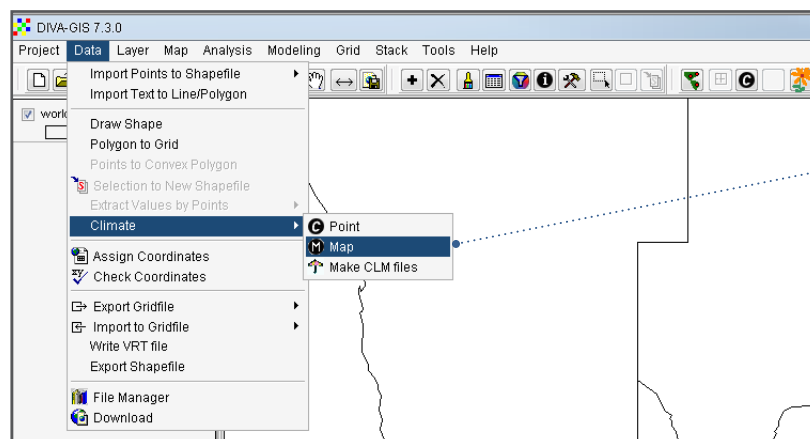
In the following analyses, you will learn how to make climate layers in the ASCII file format (*.asc) for South Africa using the 2.5 minutes CLM files (*.clm) (containing climate data) (linked to DIVA-GIS in Chapter 2).

Steps:

1. Open *World_adm0* in DIVA-GIS and zoom in on South Africa.

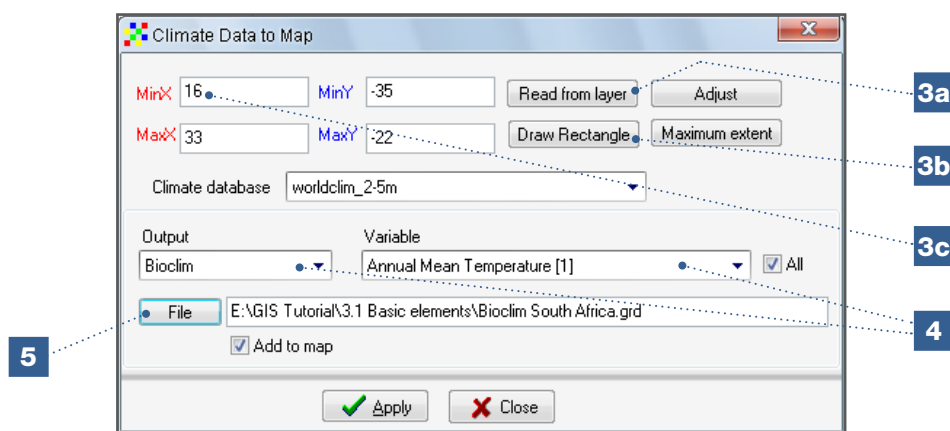


2. Go to *Data/Climate/Map*. This opens the *Climate Data to Map* window.



3. In this window, select the raster properties needed for the climate layer. This can be done by:
 - a. Selecting *Read from Layer* to use the characteristics from a raster selected in the legend of DIVA-GIS. This is useful when you wish to combine rasters and ensure they maintain the same properties.
 - b. Selecting *Draw rectangle* to define an area by drawing a rectangle in the DIVA-GIS window with the mouse.
 - c. Manually entering, defining minimum and maximum values for longitude (X) and latitude (Y) data.

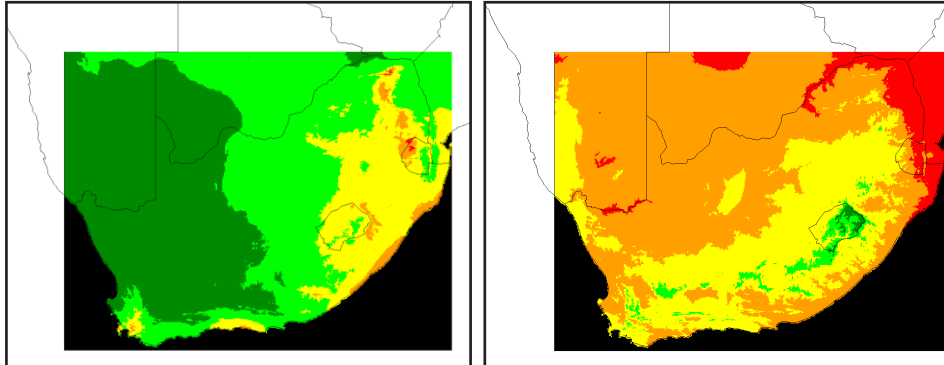
In this example, the last option is selected; entering fixed values of 16 and 33 for X and -35 and -22 for Y.



4. In addition to the definition of the raster, you will need to indicate the climate layers to be mapped, which will be given in the *Output* and *Period/Variable* dropdown menus. Options include basic temperature and precipitation data (available at monthly intervals) or the previously outlined bioclimatic variables (see Chapter 2). As the species distribution modelling in Chapter 6 will be based on these bioclimatic variables, they will also be used in this analysis. Be sure to check *All*, which will generate layers for all 19 bioclimatic variables based on an identical raster.
5. Define the output folder where the climate layers will be generated under the *File* tab.
6. In the selected folder, 19 Bioclim raster files will be created, as follows:

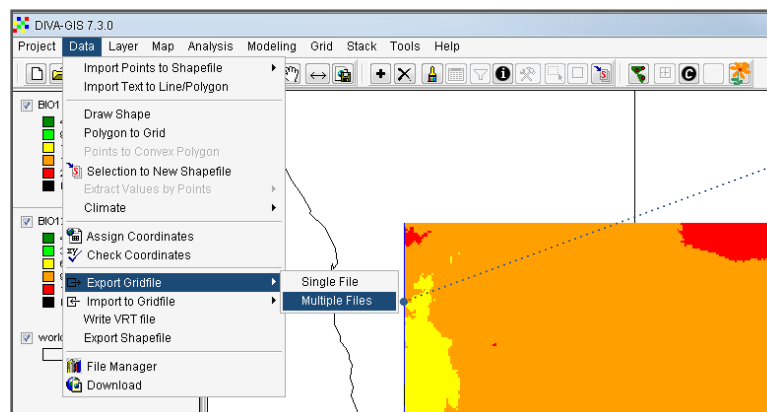
Folders	Name	Size	Type	Date Modified
	8101.grd	1 KB	GRD File	11/12/2009 12:00
	8101.gri	498 KB	GRI File	11/12/2009 12:00
	8102.grd	1 KB	GRD File	11/12/2009 12:00
	8102.gri	498 KB	GRI File	11/12/2009 12:00
	8103.grd	1 KB	GRD File	11/12/2009 12:00
	8103.gri	498 KB	GRI File	11/12/2009 12:00
	8104.grd	1 KB	GRD File	11/12/2009 12:00
	8104.gri	498 KB	GRI File	11/12/2009 12:00
	8105.grd	1 KB	GRD File	11/12/2009 12:00
	8105.gri	498 KB	GRI File	11/12/2009 12:00
	8106.grd	1 KB	GRD File	11/12/2009 12:00
	8106.gri	498 KB	GRI File	11/12/2009 12:00
	8107.grd	1 KB	GRD File	11/12/2009 12:00
	8107.gri	498 KB	GRI File	11/12/2009 12:00
	8108.grd	1 KB	GRD File	11/12/2009 12:00
	8108.gri	498 KB	GRI File	11/12/2009 12:00
	8109.grd	1 KB	GRD File	11/12/2009 12:00
	8109.gri	498 KB	GRI File	11/12/2009 12:00
	81010.grd	1 KB	GRD File	11/12/2009 12:00
	81010.gri	498 KB	GRI File	11/12/2009 12:00
	81011.grd	1 KB	GRD File	11/12/2009 12:00
	81011.gri	498 KB	GRI File	11/12/2009 12:00
	81012.grd	1 KB	GRD File	11/12/2009 12:00
	81012.gri	498 KB	GRI File	11/12/2009 12:00
	81013.grd	1 KB	GRD File	11/12/2009 12:00
	81013.gri	498 KB	GRI File	11/12/2009 12:00
	81014.grd	1 KB	GRD File	11/12/2009 12:00
	81014.gri	498 KB	GRI File	11/12/2009 12:00
	81015.grd	1 KB	GRD File	11/12/2009 12:01
	81015.gri	498 KB	GRI File	11/12/2009 12:01
	81016.grd	0 KB	GRI File	11/12/2009 12:01
	81016.gri	1 KB	GRD File	11/12/2009 12:01
	81017.grd	1 KB	GRD File	11/12/2009 12:01
	81017.gri	498 KB	GRI File	11/12/2009 12:01
	81018.grd	1 KB	GRD File	11/12/2009 12:01

- From the table in Section 2.2, derived from www.worldclim.org/bioclimate, we know that BIO1 refers to the mean annual temperature and BIO12 to annual precipitation. Add these two layers to the map.



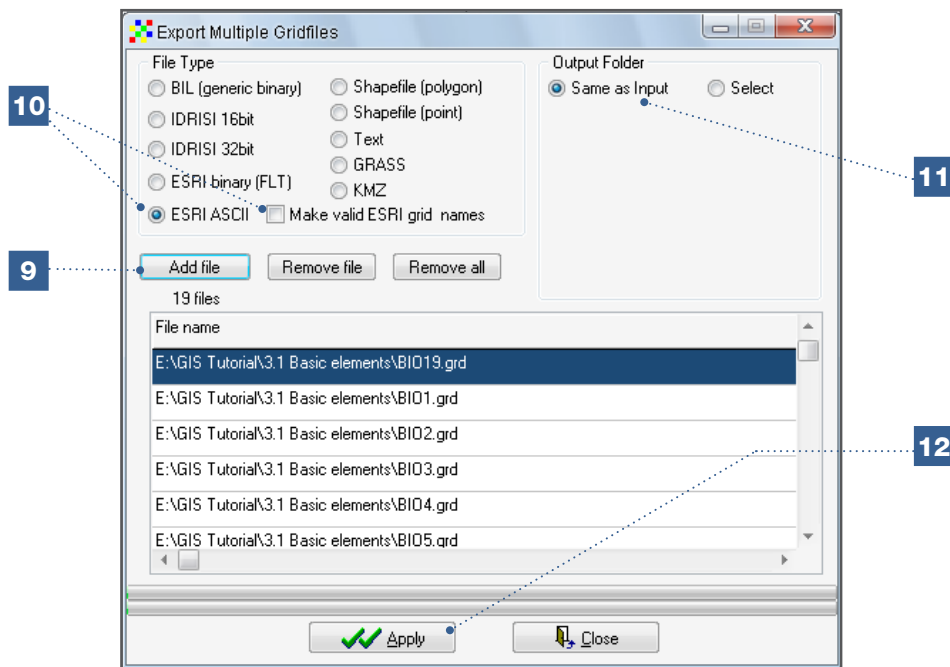
As noted in the introduction to Chapter 2, GIS and species distribution modelling programmes use a variety of raster file types. DIVA-GIS uses *.grd format while Maxent works best with rasters in ASCII format (*.asc). DIVA-GIS allows you to import and export rasters using files compatible with different programmes.

- In order to use the generated climate layers for species distribution modelling in Maxent, you must convert these to ASCII files (*.asc) by using the *Data/Export Gridfiles/Multiple Files* option.



- Select the raster files (*.grd) you wish to convert to ASCII (*.asc) using the *Add File* button.
- To export *.grd raster files as ASCII files (*.asc), select the option: *ESRI ASCII*. When exporting *.grd raster files as ASCII files (*.asc) you can choose to format the file names (by selecting the *Make valid ESRI grid names* option). For this analysis, however, do not select this option.
- Select *Output Folder* and save the newly formatted rasters in the same folder as rasters in the original file type (*Same as Input*) or in another folder (*Select*).

12. Click *Apply* to start the process.



13. The generated ASCII files (*.asc) can now be used in Maxent for species distribution modelling (see Analysis 2.2.2).

3.1.6. How to import generic climate data to DIVA-GIS

The climate layers created in the previous section (based on *Data/Climate/Map*) are derived from the CLM files (*.clm), using the 2.5-minute resolution data described in Chapter 2 for DIVA-GIS. The Worldclim website (www.worldclim.org) provides more detailed climate data (up to 30 seconds or 1 km at the equator). For analysis in a small area using highly precise presence points, these climate data might be the most appropriate; however, rasters with 30-second resolution for the entire world occupy large amounts of space on the computer. Therefore, it is useful to download climate layers with 30-second resolution in tiles of 30 x 30 degrees from the Worldclim website (<http://www.worldclim.org/tiles.php>). This section illustrates how to generate climate raster files (*.grd) from a BIL raster file (*.bil) of 30 seconds for a specific region of 30 x 30 degrees, available from Worldclim.

Future climate layers available for download from the Worldclim webpage (<http://www.worldclim.org/futdown.htm>) and from the Downscaled GCM Data Portal (<http://gisweb.ciat.cgiar.org/GCMPPage>) can be imported to DIVA-GIS in a similar manner. In Section 6.3, the use of future climate data in climate change impact studies on plant distributions and diversity is described in further detail.

Steps:

1. Go to the website: <http://www.worldclim.org/tiles.php>. Select tile 33 by clicking on it (alternatively these four files are also included in the Basic Elements folder).

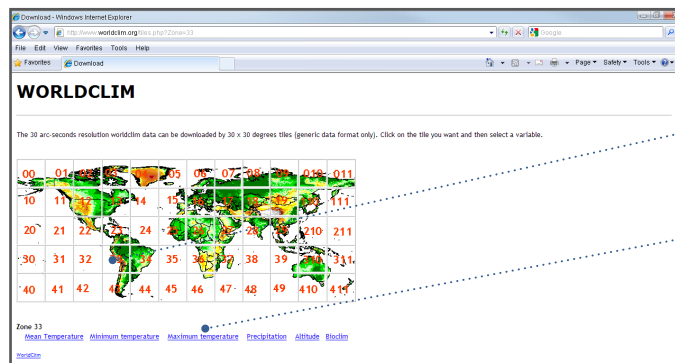
Importing rasters to ArcGIS from DIVA-GIS

Raster files created in DIVA-GIS can also be imported to ArcGIS. To do this, you must start by converting the raster files (*.grd) from DIVA-GIS to ASCII files (*.asc) by using the option *ESRI ASCII* or by converting them to FLOAT (*.fit) files, using the option *ESRI binary (FLT)*. These file can then be imported in ArcGIS.

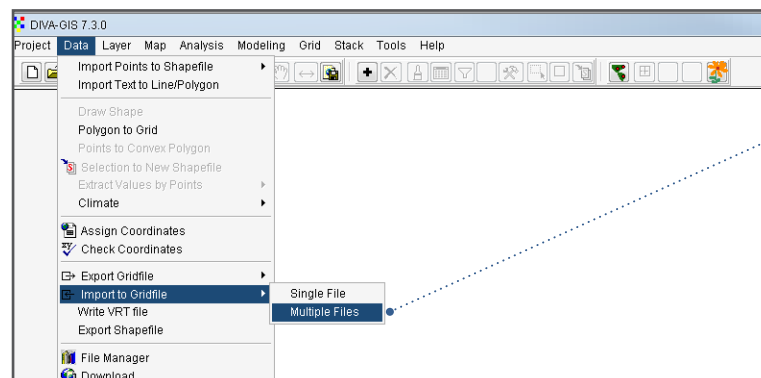
2. Download the zipped datasets for *Minimum temperature*, *Maximum temperature*, *Precipitation* (these datasets consist of monthly climate data) and *Altitude*. Save the datasets on your computer.

Note

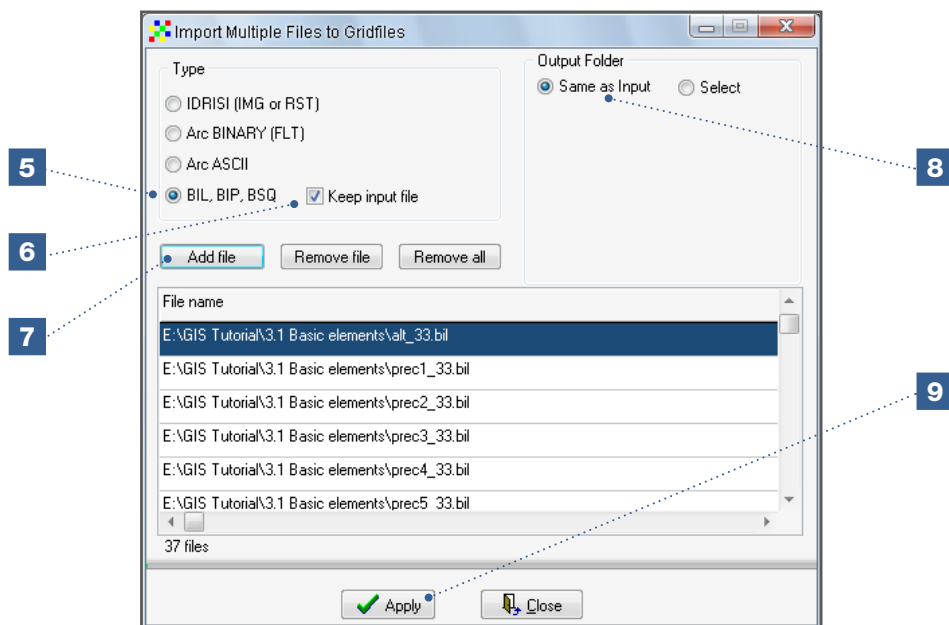
You will notice that the 19 Bioclim variables can also be downloaded. This is useful when performing species distribution modelling analyses with Maxent (see Sections 6.2, 6.3 and 6.4). For this analysis, however, only the monthly climate layers of *Minimum temperature*, *Maximum temperature*, *Precipitation* and the raster file *Altitude* are used. They are also used in Analysis 3.1.7 to create your own CLM file (*.clm).



3. Extract the layers from the zipped files.
4. In DIVA-GIS, go to *Data/Import to Gridfile/Multiple Files*.



5. Select the *BIL, BIP, BSQ* box.
6. Keep default options for the input file.
7. Click *Add file* and select the monthly climate layers of *Minimum temperature, Maximum temperature, Precipitation* and the raster *Altitude* file that you would like to convert to a *.grd file.
8. Save the generated raster files (*.grd) in the same *Output Folder* as the raster files in the original format (*Same as Input*). You may also choose to save them in a different folder (*Select*).
9. Click *Apply* to start the process.



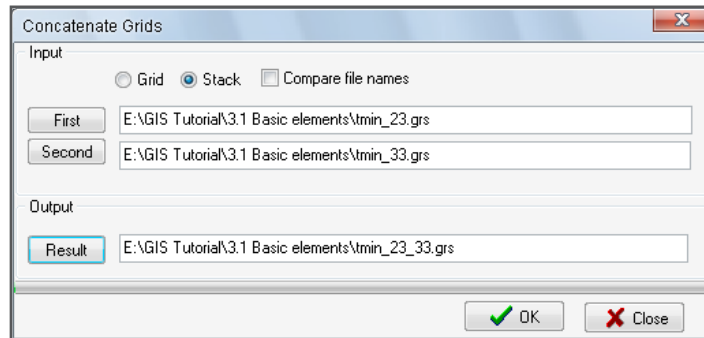
10. Now you can open the climate layers in DIVA-GIS by using the *Add layer* option.

Note

The same procedure must be followed when raster data in the following formats is being imported to DIVA-GIS: IDRISI, Arc BINARY and Arc ASCII. For example, the future climate data at the Downscaled GCM Data Portal (<http://gisweb.ciat.cgiar.org/GCMPPage>) are available in ASCII format.

Merge grids

When you want to use 30-second resolution climate layers for a specific area, the area may extend over all or part of the adjacent 30 x 30 degrees tiles. After having generated the raster files (*.grd), the climate layers of the adjacent tiles can be merged in DIVA-GIS using the *Grid/Merge* option. First, stack each dataset using the *Stack/Make Stack* option and then merge them using *Grid/Merge*.



In this image, the stacks of the 12 minimum temperature layers from zones 23 and 33 are merged.

3.1.7. How to make CLM files in DIVA-GIS

In the previous section, datasets with layers of monthly precipitation, minimum and maximum temperatures were generated. When performing an ecological analysis in DIVA-GIS, it is often easier to work with CLM files (*.clm) with its 19 bioclimatic parameters, than with the individual climate layers. For example, climate data from CLM files are used in the Analysis 3.1.4 and Sections 4.2 and 6.1. The DIVA-GIS website provides climate databases (*.clm files) of current climatic conditions up to a resolution of 2.5 minutes (5 km at the equator); see <http://www.diva-gis.org/climate>. You may, however, want to prepare CLM files with climate data of 30-second resolution, such as those generated in Analysis 3.1.6 and when using future climate data.

This section illustrates how to make CLM files in DIVA-GIS. As an example, in this analysis, CLM files will be prepared from the datasets generated in Analysis 3.1.6 (*Minimum temperature, Maximum temperature, Precipitation and Altitude*). Remember that all rasters must have the same extent and be of the same resolution, as is the case for the previously generated datasets. The preparation of CLM files in DIVA-GIS is also explained by Ramirez and Bueno-Cabrera (2009).

Steps:

1. Each dataset (*Minimum temperature, Maximum temperature, Precipitation*) contains 12 files corresponding to monthly values. Make sure the names of these files are differentiated by the numbers 1 through 12. The numbers need to be located at the end of the file names. Except for the end number, the file name should be the same in each of the 12 files. For each dataset, this is the prefix that will be used to develop the CLM file. If this is not the case, the file names need to be changed accordingly. For this analysis, the datasets of tile 33 and the names of the *.grd and the *.gri files must be changed manually.

In the case of the dataset *Precipitation*:

- *prec1_33.grd* and *prec1_33.gri* become *prec1.grd* and *prec1.gri*,
- *prec2_33.grd* and *prec2_33.gri* become *prec2.grd* and *prec2.gri*...
- *prec12_33.grd* and *prec12_33.gri* become *prec12.grd* and *prec12.gri*.

In the case of the dataset *Minimum Temperature*:

- *tmin1_33.grd* and *tmin1_33.gri* become *tmin1.grd* and *tmin1.gri*,
- *tmin2_33.grd* and *tmin2_33.gri* become *tmin2.grd* and *tmin2.gri*...
- *tmin12_33.grd* and *tmin12_33.gri* become *tmin12.grd* and *tmin12.gri*.

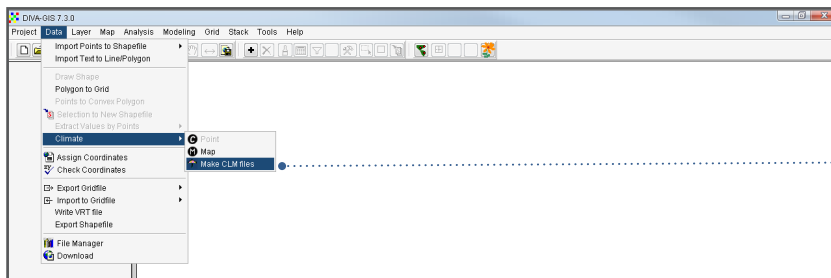
In the case of the dataset *Maximum Temperature*:

- *tmax1_33.grd* and *tmax1_33.gri* become *tmax1.grd* and *tmax1.gri*,
- *tmax2_33.grd* and *tmax2_33.gri* become *tmax2.grd* and *tmax2.gri*...
- *tmax12_33.grd* and *tmax12_33.gri* become *tmax12.grd* and *tmax12.gri*.

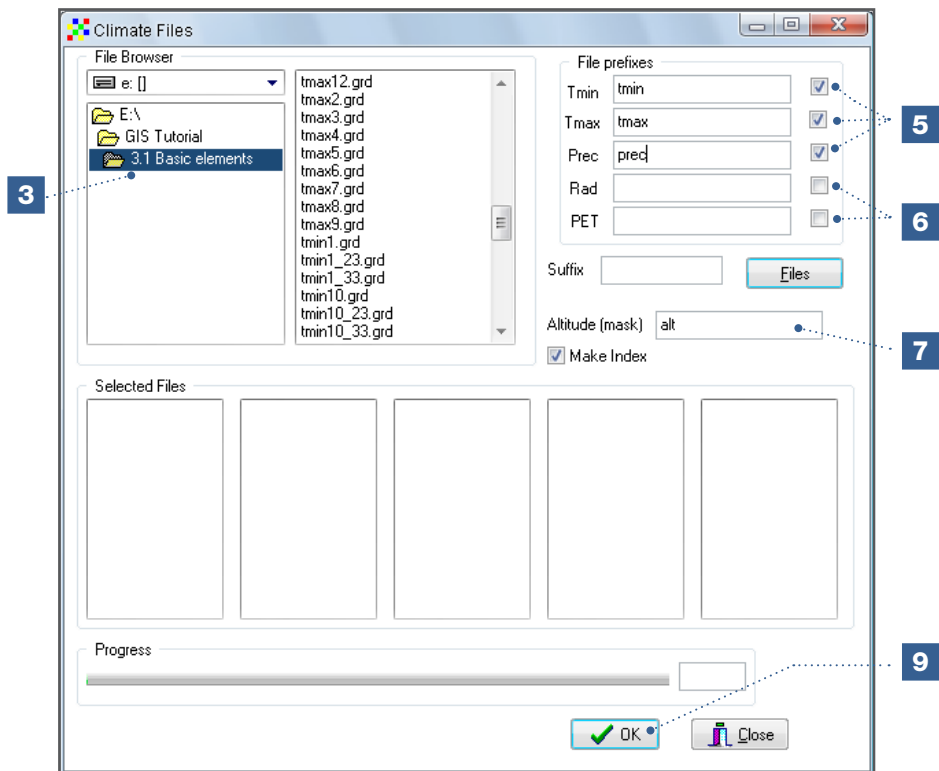
Note

Renaming of the 72 files might be facilitated by using batch rename software (e.g. Rename Master).

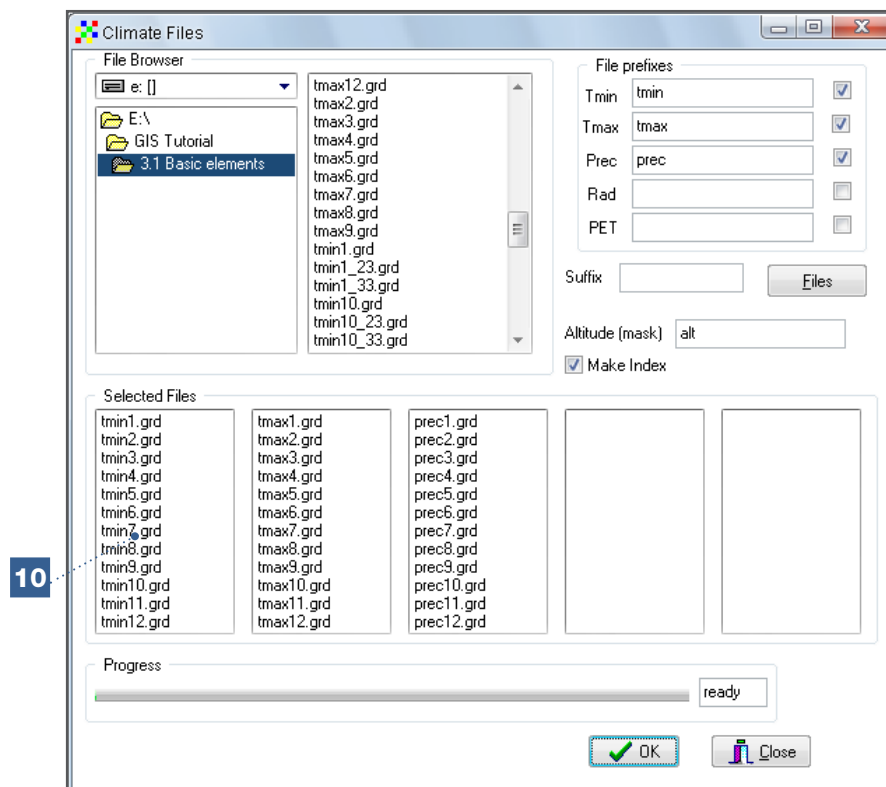
2. After having changed the file names, go to *Data/Climate/Make CLM files*.



3. Under *File Browser*, select the folder where the *Minimum temperature*, *Maximum temperature*, *Precipitation* raster files (*.grd) and *Altitude* raster were saved.
4. All raster files (*.grd) in the indicated folder are shown at the right hand of the *File Browser*.
5. Indicate the raster file prefixes in the boxes *T min* (minimum temperature), *T max* (maximum temperature) and *Prec* (precipitation). The prefixes include all the characters before the number (1-12). For this analysis, the prefix of the *T min* is *tmin*; the prefix of *T max* is *tmax*; and the prefix of *Prec* is *prec*. Make sure the boxes to the right of the selected parameters are checked.
6. There is also an option to include data of radiation (rad) and potential evapotranspiration (PET). Such data are not necessary for the 19 bioclimatic variables and this data is currently not available at Worldclim. Therefore, they will not be used in this analysis.
7. Indicate the altitude (alt) layer.
8. Keep all other options as default.
9. Click *OK* to start the process.



10. After having clicked OK, the layers of the three datasets should be imported automatically in the first three columns of the selected files.



Note

The process of making CLM files (*.clm) can be demanding if you have a computer with relatively low processing capacity. If this is the case, it is better to use climate layers with a lower resolution (5 or 10 minutes).

The CLM files (*.clm) are saved as the following:

- *climate.cli*
- *alt.clm*
- *index.clm*
- *tmin.clm*
- *tmax.clm*
- *prec.clm*

The (*.cli) file contains information about the characteristics of the CLM files (*.clm). It is recommended to change the name of the *climate.cli* file to a more specific name in order to be able to distinguish the CLM files (*.clm) from other CLM climate databases. In this analysis, for example, the name of the *climate.cli* file can be changed in *wclim_t33_30sec.cli*. Do not change the names of the CLM files (*.clm); otherwise they will not be recognized in DIVA-GIS.

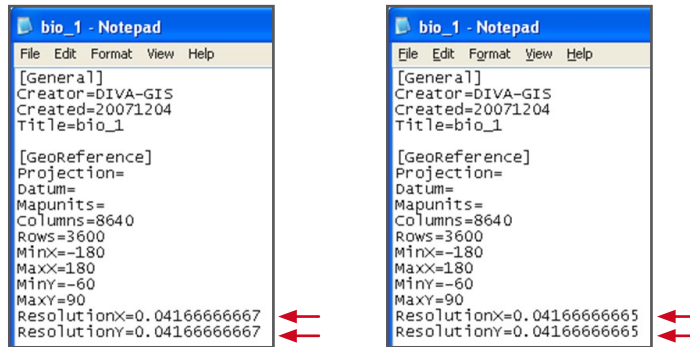
Assuring identical raster properties before combining rasters

Sometimes an error occurs when two or more rasters are combined during calculations in DIVA-GIS (e.g. overlay) or when two stacks of rasters are used as input in Maxent (e.g. to compare areas of potential distribution of a species under different environmental scenarios or climates) (see Section 6.3). This is usually due to differences in raster properties.

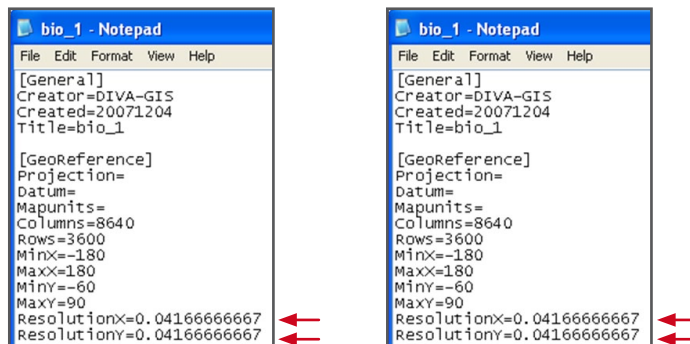
Raster properties, i.e. resolution (cell size), extent [number of rows and columns, raster corners or vertices (min and max X and Y values)], need to be identical in order to combine them in DIVA-GIS, as explained in Analysis 3.1.3, and to use them as inputs in Maxent after being converted to ASCII format (*.asc). To accomplish this in DIVA-GIS, rasters should be created with identical properties, as mentioned in Analysis 3.1.5.

Nevertheless, after using these options to create rasters of identical properties, very small differences in the decimals of the resolution and/or in the coordinates may still remain. This may happen if rasters of identical properties are created from different datasets imported to DIVA-GIS (e.g. current and future climate data sets or soil and climate datasets).

Therefore, if an error occurs in the calculations when combining rasters or stacks in DIVA-GIS or in species distribution modelling with Maxent, it is recommended to verify the differences in resolution and/or in coordinates for the vertices of the rasters. This can be done by opening the raster files (*.grd) using *Notepad*. Look at the example below and note the information in the raster (*.grd) for *Mean Annual Temperature* (BIO1) under the current climate and future climatic conditions in Southeast Asia.



Even though the number of columns and rows and the coordinates for the vertices (*MinX*, *MaxX*, *MinY*, *MaxY*) of both rasters are the same, there is a difference in the eleventh decimal of the resolution (*Resolution X* and *Resolution Y*). The difference is extremely small and will not lead to variations in the visualization but will generate an error preventing the comparison of potential distribution areas for current and future climates when converting rasters to ASCII format, appropriate for Maxent. To solve this problem, the different values can be manually changed in the raster documentation files (*.grd) to make them equivalent.



After adjusting the eleventh decimal, the resolutions are now identical and the rasters can be combined. This solution is appropriate for adjusting any kind of minor difference which exists in the coordinates of the vertices (*MinX*, *MaxX*, *MinY*, *MaxY*).

3.2. Exporting layers to Google Earth

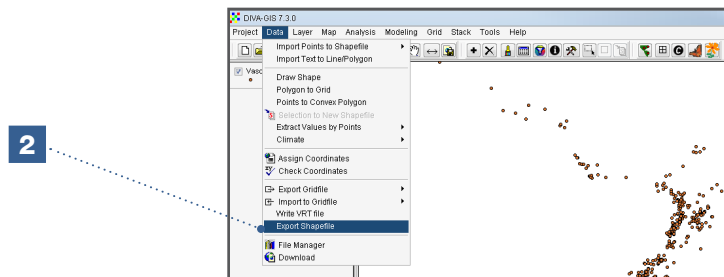
Google Earth uses specially formatted files: *.kml and *.kmz. These formats are exclusive to Google Earth; therefore, information (layers) must be converted to files using this format in order for the data to be visualized using Google Earth. DIVA-GIS Version 7.1 includes the option for exporting both vector and raster layers to Google Earth.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Google Earth 	<p>Data Files:</p> <p>Folder 3.2 Export to Google Earth</p> <ul style="list-style-type: none"> • <i>Vasconcellea species</i> (shp, shx, dbf) • <i>Distribution Pinus kesiya</i> (grd, grf)

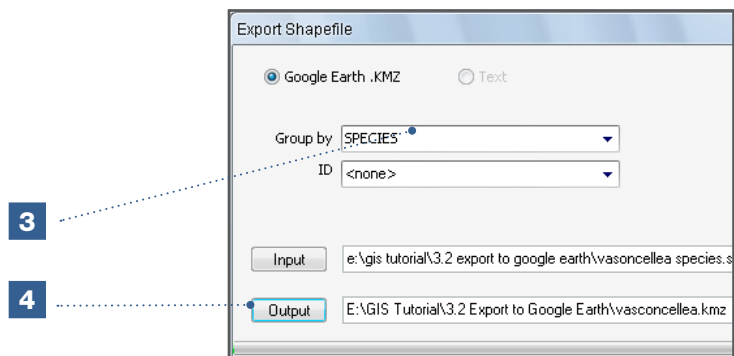
3.2.1. How to export data to Google Earth

Steps:

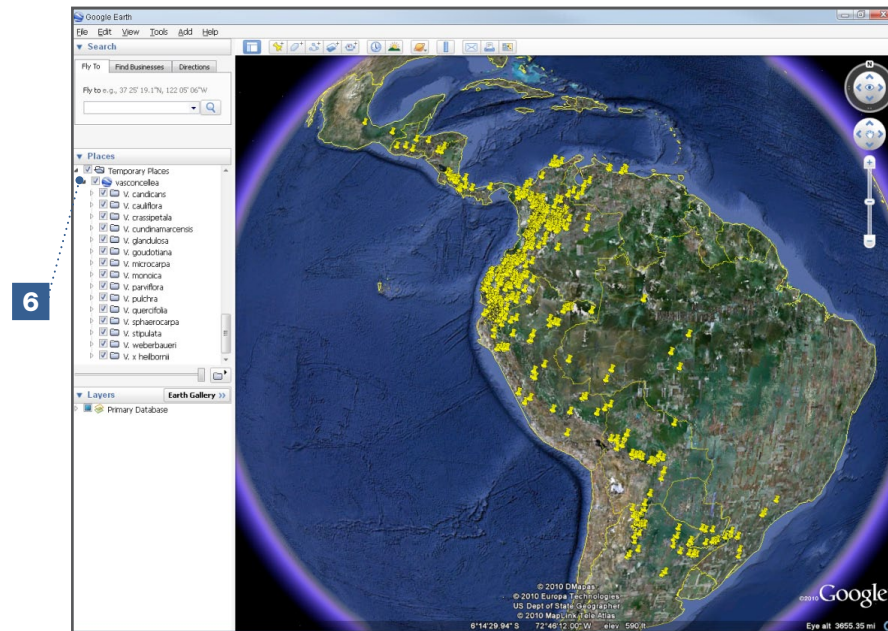
1. To export a vector file (*.shp) to Google Earth, it must first be opened with DIVA-GIS. For this analysis, open the *Vasconcellea species.shp* file.
2. Select *Data/Export Shapefile*.



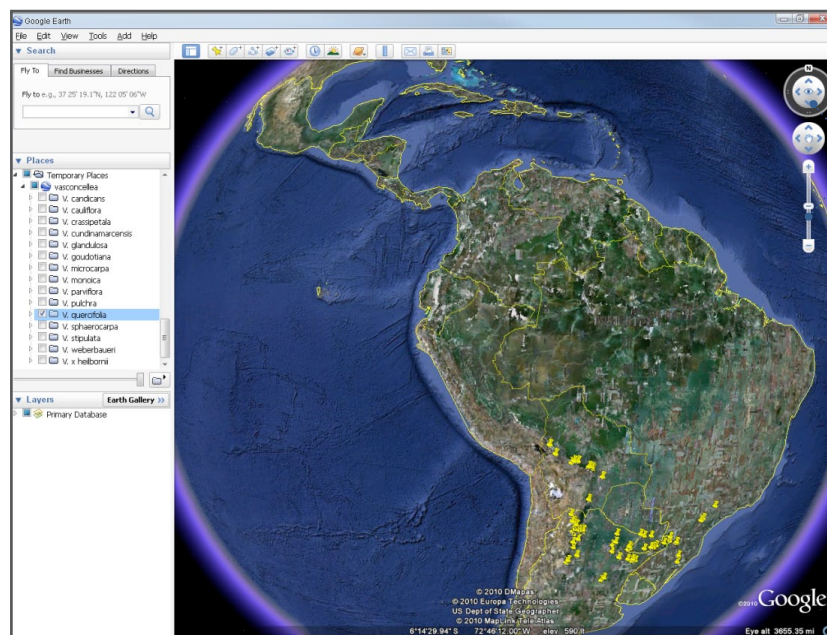
3. In the *Export Shapefile* window, under the *Group by* option, select *Species*. This will allow the visualization of a specific species in Google Earth. The *ID* option allows you to add text to each point. Since the *Vasconcellea species* file has multiple points, for this analysis it is recommended to select the option: <none>.
4. The *Output* button allows you to name the newly generated *.kmz file.



5. Go to the folder where the *.kmz file was saved and open the generated file.
6. In the *Places* window you will find the folder, *Temporary Places*, containing the newly generated file. Click on the triangle sign and notice that *.kmz file contains subgroups. Select only the *V. quercifolia* species subgroup.



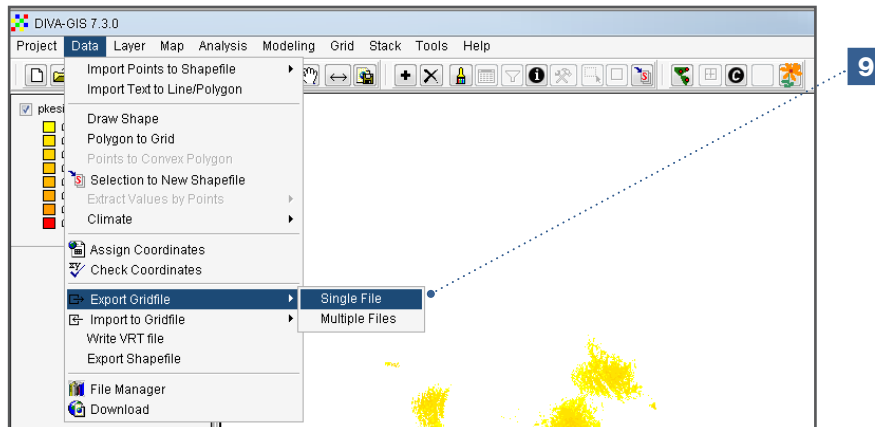
7. Using the *Pan* and *Zoom* tools in Google Earth, locate the area in northern Argentina where there is a large presence of *Vasconcellea quercifolia* (around the city of Salta). Notice that some points are located in downtown Salta, which illustrates the danger of using a very high resolution. Most likely, it was not possible to georeference these points to the same degree of precision or resolution as the satellite imagery that can be viewed in Google Earth.



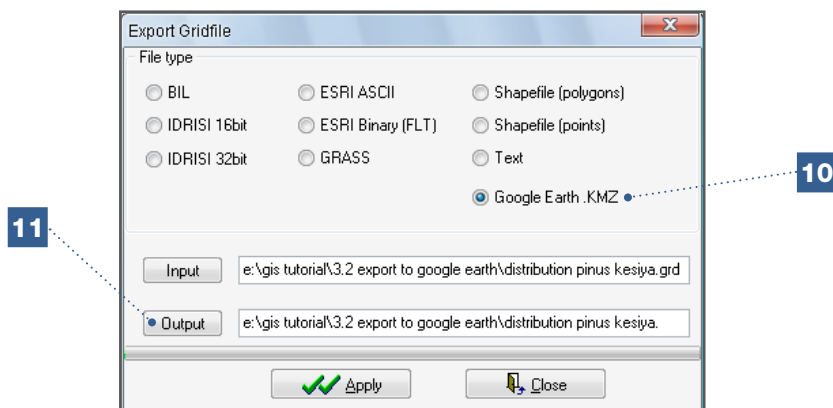


The next activity consists of visualizing a raster, such as those resulting from a diversity analysis (see Chapter 5) or species distribution modelling (see Chapter 6). In this analysis, the species distribution modelling of an Asian pine tree, *Pinus kesiya*, is used as an example.

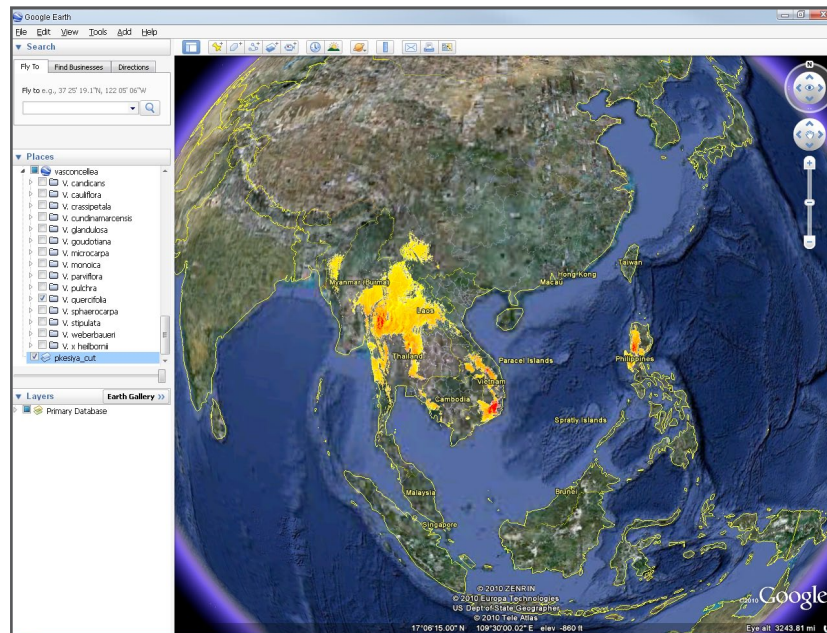
8. Open the file: *Distribution Pinus kesiya.grd* in DIVA-GIS.
9. In the menu, go to *Data/Export Gridfile/Single File*.



10. The *Export Gridfile* window will display automatically; select *Google Earth.KMZ*.
11. Use the *Output* button to define the location and name of the *.kmz file.

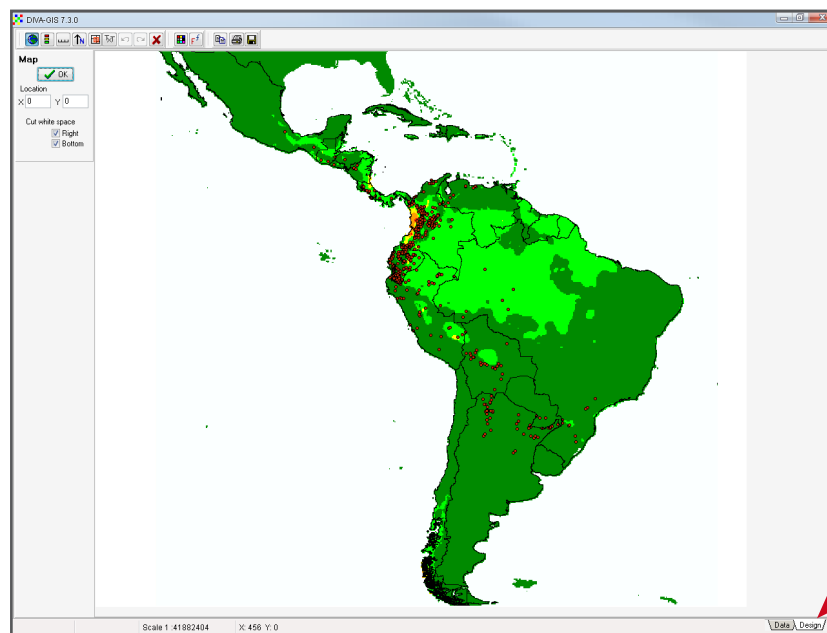


12. Finally, go to the folder where the *.kmz file is saved and open it.





3.3. Editing maps and finalizing a project

If you wish to display maps in documents or reports, you can customize the combination of generated layers, under the *Design* option, located in the lower right-hand corner of the screen. Selecting this option allows you to carry out basic editing of the current visualization (e.g. adding scale, adding north arrow, adding text).



After customizing the map, the image can be saved as a *.png, *.tif or *.bmp file, or can be copied directly to the clipboard. The *Map* menu, under the *Data* tab, includes the *Map to Image* option, which allows you to save the edited map as a *.png, *.tif or *.bmp file, but does not allow for any further editing.

Saving projects in DIVA-GIS

Vector and raster files are saved automatically when generated and modified in DIVA-GIS. A specific combination of files can be saved as a project, which is defined as a combination of different raster and vector files. If you wish to return to your work, simply re-open the project; this will automatically open all layers that constitute the map, so you will not need to create the map again. Use the  icon or select *Project/Save (as)* to save a project. To open a project saved at an earlier time, use the *Project/Open* option or click on the  icon.

A project will be saved as a *.div file. This file only contains the location on the hard disk (path) of each layer in the project. Therefore, it can only be opened on the same computer on which it was originally created (provided there are no changes in the location of the layers). To share or exchange projects, go to *Project/Export Project*. Before exporting, convert the *.div file into a *.dix file, which is interchangeable. Remember that a *.dix file can be very large, depending on its contents (resolution and number of layers), which makes it difficult to exchange. Individual vector and raster files can also be shared, but be sure to share all files which make up a single layer:

- Vector files: *.shp, *.shx and *.dbf
- Raster files: *.grd and *.gri (the image files *.bpw and *.bmp are generated automatically when visualizing a raster file but are not an essential part of the file).

References

DIVA-GIS. 2005. User Manual, version 5.2 [on line]. Available from: http://www.diva-gis.org/docs/DIVA-GIS5_manual.pdf. Date accessed: October 2010.

Ramirez J, Bueno-Cabrera A. 2009. Working with climate data and niche modelling. Creation of bioclimatic variables. Tutorial [on line]. Available from: http://gisweb.ciat.cgiar.org/GCMPPage/docs/tutorial_bcvvars_creation.pdf. Date accessed: October 2010.

Chapter 4

Quality control

One of the main objectives of undertaking the spatial analysis of biodiversity data is to provide information to assist in effective policy- and decision-making processes for natural resource conservation and use. In order to formulate appropriate management and conservation strategies, it is critical that datasets are of high quality and are precise (Chapman 2005a, 2005b). The use of incorrect or low-quality information may have significant consequences on the relevance and appropriateness of subsequent recommendations, decisions and even investments. Genebank and herbarium data (available through biodiversity networks such as the GBIF (see Section 2.3) are increasingly used in biogeographical studies; however, such data are from third parties and the origins are often unknown, making the issue of data quality even more pertinent.

The specific objective of this chapter is to show users how to identify possible erroneous presence points using different tools and how to take corrective actions to ensure high levels of data quality. Poor data quality can result from various causes, such as errors in site descriptions, imprecise coordinates or even mistakes or changes in taxonomic identification. Errors are frequently made when recording coordinates in the field, especially when a data transcription step is included, or when entering data into a database. Georeferencing records from an office setting or at a distance, based on site descriptions, can also lead to a poor data quality.

Two key aspects of data quality include the accuracy and precision of geographic coordinates. The accuracy of coordinates determines the ability to correctly represent the site of collection/observation of a presence point. Precision refers to the level of detail of the coordinates necessary to represent the described site effectively. Precision can be assessed by reviewing the method in which the coordinates were determined (e.g. maps versus GPS) or according to the number of decimals included in the coordinates, as already discussed in Section 2.1. A lack of accuracy in the analyzed data will inevitably lead to errors in the output/results of the analysis, while a lack of precision will often result in conclusions of limited use, as they are only representative at a very low resolution. Data can be very precise but inaccurate and can also be very accurate but highly imprecise.

DIVA-GIS can help in making decisions regarding correction or elimination of presence points displaying quality problems. The software has some useful tools to identify possible errors in coordinates based on the existing administrative unit information in the passport data (see Section 4.1) or to identify suspicious presence points based on atypical environmental conditions, which can indicate a taxonomic misidentification or erroneous coordinates (see Section 4.2).

Another important aspect of data quality, and one which is difficult to evaluate, is bias. Bias generally occurs when a sample is not wholly representative of the area being studied. This can be effectively remedied with a sound data collection strategy. Nevertheless, many spatial biodiversity analyses are made with some or all data originating from herbaria and genebanks. Such data are usually not generated for the purpose of biogeographical studies and often entail *ad hoc* collecting, non-systematic sampling and uneven sampling efforts (Chapman 2005a). Frequently, specimens have been collected from easily accessible areas or areas where a species is known to occur, thus negatively affecting the representativeness of the data (Hijmans et al. 2000). These

issues can lead to a sample population which may or may not be representative for the species in terms of a(n) environmental or geographical space, as the data provide information on patterns found only at the sampled sites rather than across the entire study area (Williams et al. 2002). Several methods exist to reduce the sample bias of a dataset, but these can only resolve the problem to a limited extent. Chapters 5 and 6 discuss some of these methods in further detail (e.g. circular neighbourhood, rarefaction and species distribution modelling).

4.1. Quality control based on administrative unit information

One way to evaluate the accuracy of a presence point is to compare the administrative unit data included in its passport data with administrative unit information extracted from thematic layers based on the geographic coordinates of the point. In order to do this, the passport data needs to include data for the country and, preferably, for lower administrative units. The *Data* menu of DIVA-GIS provides the *Check Coordinates* option, which allows one to check the quality of coordinates based on the administrative unit data.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel 	<p>Data Files:</p> <p>Folder 4.1 Quality control – Administrative units</p> <ul style="list-style-type: none"> • <i>Vasconcellea final errors</i> (shp, shx, dbf) • <i>Latin America countries</i> (shp, shx, dbf) • <i>Latin America Adm 01</i> (shp, shx, dbf)¹

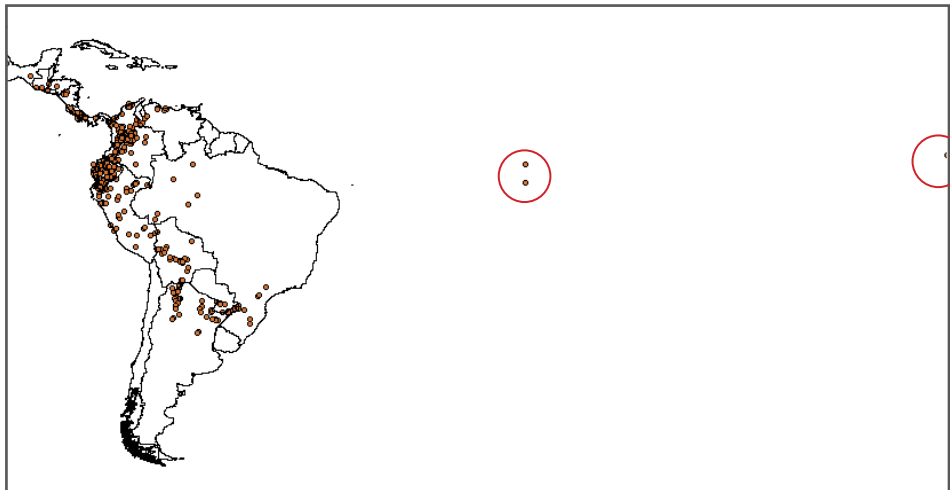
4.1.1. How to verify data quality based on passport administrative unit data

In the following analysis you will use a portion of the dataset corresponding to information on the geographic distribution of highland papayas; this information was collected during a highland papaya study conducted in Latin America (Scheldeman et al. 2007). In view of the objectives of this analysis, some errors have been introduced into the dataset. The clean version of this dataset will be used in the inter-specific diversity analysis of Section 5.1.

¹ Note that the Latin America Adm 01 layer contains some imprecisions at administrative level 1. These do not influence the analyses in this exercise but might prevent this file being used in other analyses. For an up to date maps at Adm 01 level please visit the DIVA-GIS website (<http://www.diva-gis.org/Data>) or the GADM database of Global Administrative Areas (<http://www.gadm.org/>).

Steps to use the check coordinates function in the data menu of DIVA-GIS:

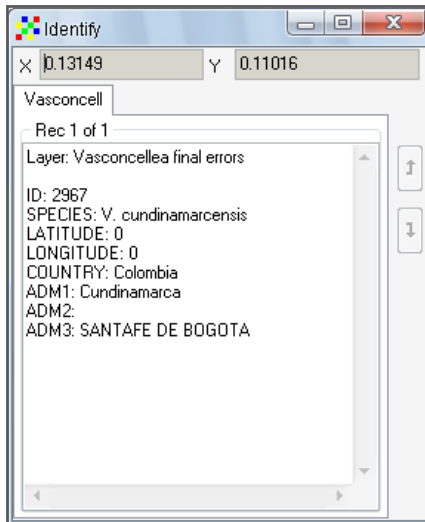
1. The first way to identify potentially erroneous points is to visualize the dataset on a map. Open the *Vasconcellea final errors.shp* file to visualize the *Vasconcellea* collection points in Latin America. To see all points on the map use the *Zoom to Theme* button, which displays the full extent of all open datasets.
2. Now, add the polygon file with the Latin American countries (*Latin America countries.shp*). Data errors are immediately obvious as some points are located outside the study area (Latin America). To access the passport data for each point and determine the ID use the *Information* button (i) in the main menu.



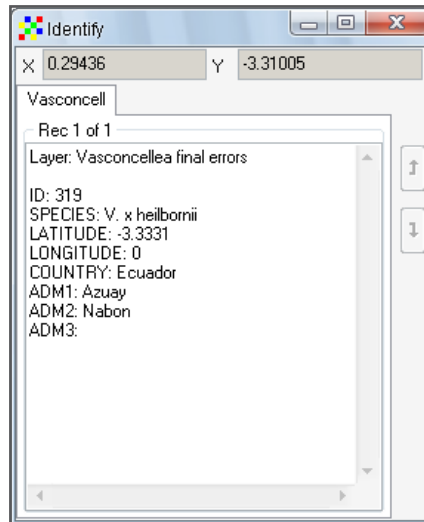
For the purpose of this analysis, the data presented illustrates three typical errors, all apparent due to the location of points outside the study area:

- **Case A: ID 2967.** The point's coordinates have values of 0,0 (which locates this point in front of the coast of Africa). This error can occur in the database if there is an absence of data, as some programmes automatically fill empty cells with a value of zero (0).
- **Case B: ID 319.** The situation is similar to the previous one, except in this case incorrect information is only present for one coordinate (longitude). In a database, it is possible for one geographic coordinate of a presence point to be missing. A question to consider: What would happen if the value for latitude were zero (0) instead of for longitude? Would the error also be as obvious and easy to identify?
- **Case C: ID 1669.** This point is far away from the study area. Here, the error has likely occurred due to the omitting of the negative sign for points located in southern latitudes or western longitudes, resulting in incorrect positions on the map.

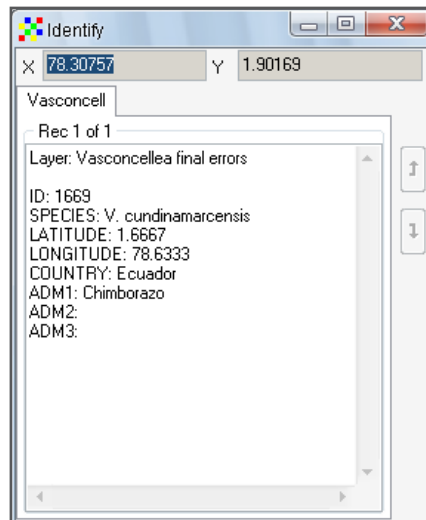
Case A



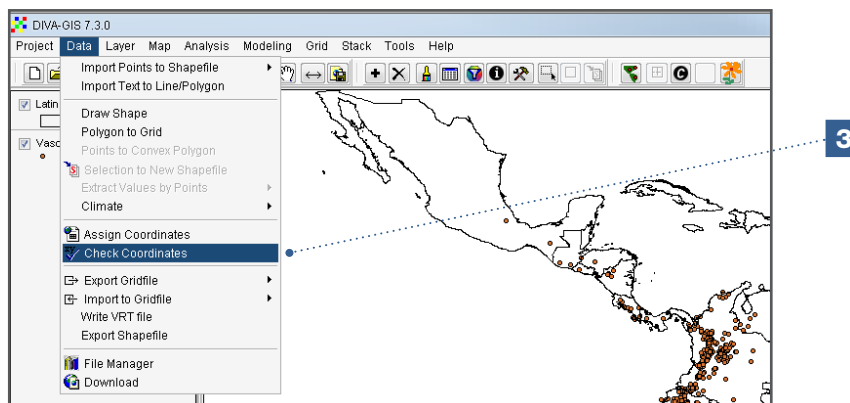
Case B



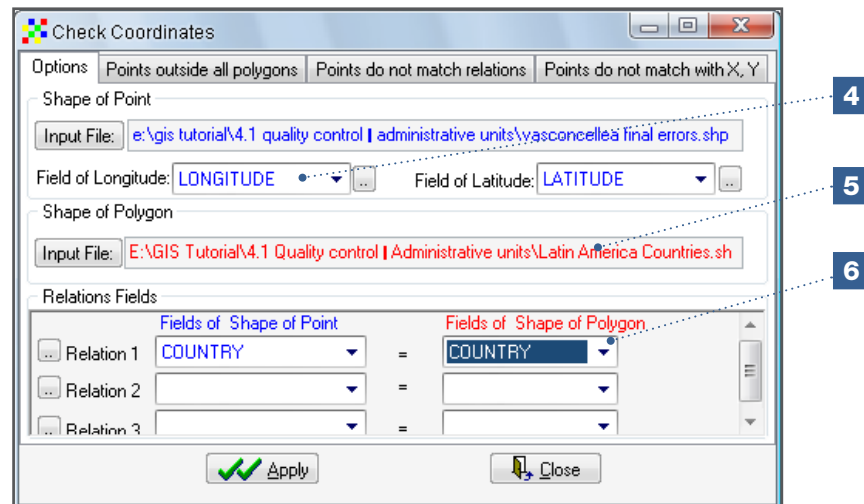
Case C



- While these errors are easily detected through the visualization of the map, the whole database must also be checked for less evident errors. To do this, go to the *Data* menu and select *Check Coordinates*. If the layer with the points has been selected, this layer will become the *Input*; otherwise click on *Input File* and select the layer with the observations.



4. Select the field with the data for longitude and latitude.
5. Select the polygon file with the administrative unit data (*Latin America countries.shp*) starting with the higher level data (the layer with countries).
6. Indicate the relationship between both files (*Country-Country*) and click *Apply*. DIVA-GIS will immediately check the inconsistencies in the information in both layers.



7. The first group of inconsistencies includes those points outside all the polygons (*Points outside all polygons* tab). In this analysis, these points are located outside the polygon of Latin America. This option is very useful for identifying points with unlikely locations, such as a point of a plant species located in the ocean.

No. Reg	X	Y	ID	SPECIES	LATITUDE	LONGITUDE	COL
6	0	-3.3331	319	V. x heilbornii	-3.3331	0	Ecu.
321	-80.8419	-1.5667	796	V. parviflora	-1.5667	-80.8419	Ecu.
561	-80.6833	0	1367	V. parviflora	0	-80.6833	Ecu.
622	78.6333	1.6667	1669	V. cundinam	1.6667	78.6333	Ecu.
830	0	0	2967	V. cundinam	0	0	Colo.

In the fourth column of the information displayed, check the identification (ID) of data with errors. With this information, you can return to the original data file [in Excel or in a dBase IV file (*.dbf)] and check the details in order to determine how to correct the error. The options *Pan to* and *Zoom to* (see Chapter 3) also allow you to identify other points that may include mistakes. These problem points can also be exported to a text file using the *Export* function.

Note

It is strongly recommended to document, in detail, all changes made to the database, including the error reports, and to keep the original database untouched, making new and corrected versions for further processing.

The list generated by DIVA-GIS in Step 7 reports five points with potential errors, three of which were already detected (ID: 2967, 319, and 1669) on the map. The other two suspicious points are:

- **ID 1367**, which is located 500 km away from the coast of Ecuador. The level of *Zoom* used in the analysis did not allow for the point to be detected at first sight. This point has a zero (0) latitude value, as well, which probably explains this error.
- **ID 796**, which is located in the ocean, about 1 km away from the Ecuadorian coast. Considering that the point's passport data indicate a collection site very close to the coast, the error is most likely due to a slight imprecision in the GPS or in the manual process of georeferencing the site. This may be corrected by assigning coordinates located nearby at the coast of an island or mainland.

What to do with erroneous points?

There are three options to deal with errors: correct, delete or keep. The most appropriate action depends on several factors. Two important factors are the scale of the study (global, continental, national or local) and the number of points available for the studied taxon. In either case, the first step is to locate the original information and verify the origin of the point (thus, the importance of maintaining field books). This step may allow you to correct the mistake immediately. If the original information is not available (which is often the case), the next step is to identify the possible cause of the error (refer to the example above for the point with ID 1669, where the only problem was the missing negative sign). If these two steps are not possible, you should consider eliminating the point. Before doing so, however, consider the consequences of such an elimination on the results (ensure that you are not creating bias). You may want to keep the erroneous data if you are working at a very large scale and the problem point is considered to be important (e.g. in a study to know the number of species per country, the exact location within a country is less important).

8. Now, look at the other errors. In the *Check Coordinates* menu, go to the *Points do not match relations* tab. The result of comparing passport data against its effective location on the map is displayed, revealing the points where passport country information does not coincide with the country location of the point on the map.

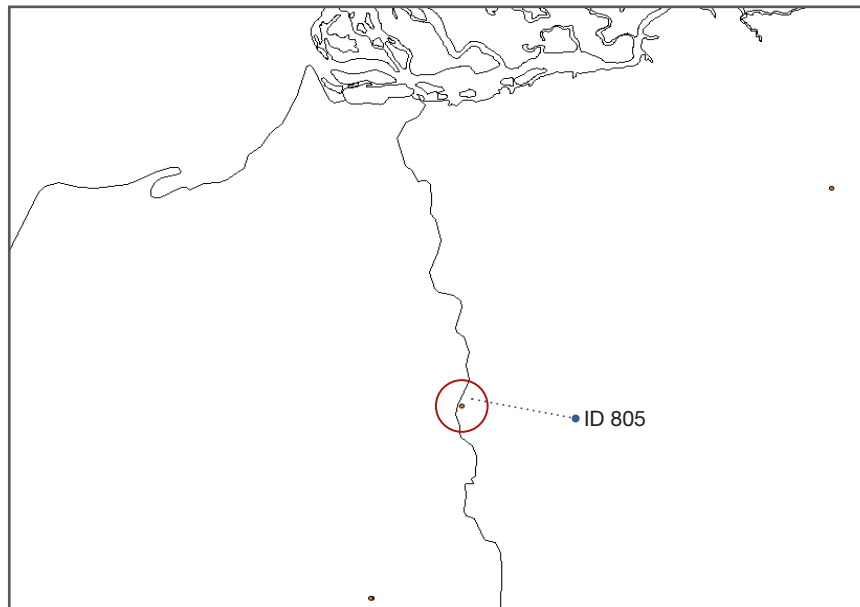
No Reg	X	Y	Point:COUN	Polyg:COUN	ID	SPECIES	LAT
326	-80.1915	-3.6833	Peru	ECUADOR	805	V. parviflora	-3.68
483	-72.7333	8.3	Brasil	COLOMBIA	1143	V. microcarpa	8.3
695	-65.7333	-2.6166	Brazil	BRASIL	2724	V. microcarpa	-2.6166

Row: 0 of 0

Buttons: Highlight, Pan To, Zoom To, Export

This analysis allows you to identify further errors, in addition to those detected after the first analysis described in Step 7:

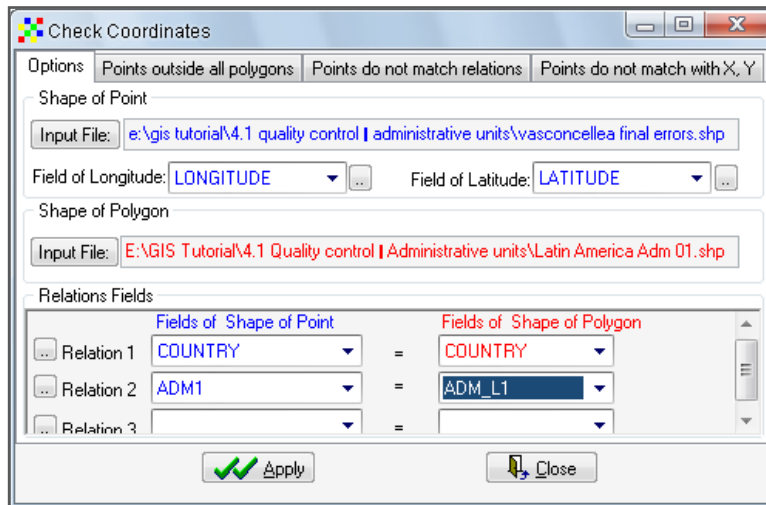
- **ID 805:** According to its passport data, this sample was collected in Peru, but according to the map, the point is located in Ecuador. Since the imprecision in the coordinate is less than 1 km, the mistake may indicate an error generated when georeferencing or recording the coordinates. The collector may have crossed the country boundary unknowingly, which could be the origin of the inconsistency in the administration data. This is a common problem and the decision as to what to do with the data point is difficult, as it may well be that an error in the descriptive part of the passport data is causing the inconsistency, while the coordinates are correct. Many studies allow a margin of error around the borders of administrative unit data, for example 5 km.



- **ID 1143:** The point is located in Colombia, while its passport information indicates it was collected in Brazil. Considering there is a huge gap between the site and the Brazilian border, this is clearly a significant error which needs to be resolved (if the cause of the problem can be identified) or the point must be deleted. In this example, the problem is a missing negative (-) sign in one of the coordinates.
- **ID 2724:** This point also illustrates a common problem, which is caused by mistakes or discrepancies in the spelling of administrative units (i.e. 'Brazil' versus 'Brasil'). This is obviously not an erroneous point, but rather a difference in spelling/language and can be easily corrected, though it is not always necessary to do so.

Data control at lower level administrative units

When passport data includes information at a lower administrative level, e.g. departments or provinces, it is recommended to run a check for errors at this level, as well. The initial procedure is the same as explained in Steps 1 to 5, above. In Step 6, however, where relationships are indicated, the layer with the lower administrative unit data (*Latin America Adm 01.shp*) must be added. After doing so, continue with Steps 7 and 8, simply adding one more level in the relationship field with the *Fields of Shape of Point* and *Fields of Shape of Polygon* tools.



New presence points with possible erroneous coordinates are apparent when double-checking this second level of data. In addition to those points previously described (ID 805, ID 1143, ID 2724), which showed errors at the country level, and the points (ID 319, ID 796, ID 1367, ID 1669 and ID 2967), located outside of the study area, these error points include:

- **ID 689:** This point is located on the boundary between two provinces in Ecuador (Morona Santiago and Chimborazo). As was the case with ID 805, the problem may be related to an imprecise coordinate, but the most probable explanation is that the collector crossed the provincial boundary unknowingly, resulting in an error in the descriptive data. When the distance between the point and the boundary is small (as is the case here), generally there is no need to take corrective action.
- **ID 2729:** The latitude of this point is listed as zero (0) - a mistake that was previously described (see point, ID 319). This point clearly illustrates the need to establish controls at lower administrative levels, as points with this type of error are not always located in the ocean or in another implausible site.
- **ID 2870:** The collection site has different names at the Colombian department level. When the point was recorded, the department's site name was 'Cundinamarca', but several zones in this locality (including the collection zone) were later renamed as the 'Distrito Capital'. The change of a name in the administrative unit often generates errors when using historic data from herbariums or museums.
- **ID 2906:** This point has a spelling mistake - a common error when using characters such as the 'ñ' which do not exist in all languages.
- **ID 2943:** Another instance of a zero (0) value input for the latitude.
- **ID 3023:** A problem in the name of a department: 'Valle' (abbreviated, common name) versus 'Valle del Cauca' (complete name).
- **ID 3068:** This point has the same problem described for the point with ID 2729 above.

4.2. Quality control through the identification of atypical points

Atypical points, or outliers, are presence points located outside the limits of the species' normal environmental ranges. Atypical data occupy an ambivalent place in spatial biodiversity analyses. Outlier identification methods can help to detect erroneous presence points in a dataset, which should be removed to ensure data quality. An atypical environment might be an indication of incorrect presence data resulting from different types of errors: erroneous geographical coordinates, erroneous taxonomic classification or introduction of individuals in places which do not correspond to their range of natural occurrence (e.g. production systems and botanical gardens).

On the other hand, an atypical point can also indicate an individual or a group of individuals that have adapted to environmental factors different to those of the most naturally occurring individuals and populations. The individual or group of individuals occurring at the location of an atypical point may have unique characteristics and provide interesting genetic material for breeding programmes interested in adaptive resistance to environmental stresses such as drought or extreme temperatures. Atypical points may also indicate a sample bias, such as under-sampling in the geographic areas/environments where the atypical points are found.

Although the different explanations for atypical points mentioned above (error, individuals growing in an extreme climate, sample bias) are all relevant in spatial biodiversity analysis, it is often difficult to determine the actual cause for such points. Consequently it is hard to decide what to do with them, i.e. to eliminate them from or include them in the dataset. The first step to determine appropriate corrective action is to locate the original information and verify the origin of the points. However, if the data was sourced from third parties (like the GBIF), this is often not possible and the decision to eliminate or keep atypical points requires careful consideration.

Points are likely to be erroneous when they reflect a completely different climate than the rest of the dataset; if this is the case, these points should be removed. It is more difficult to determine a threshold from which a point can be considered an outlier, with sufficient probability that it is erroneous. Various statistical methods (uni- and multivariate) are available for identifying such points (see Chapman 2005a). This section outlines two methods included in DIVA-GIS to detect outliers:

- *Reverse jackknife* (Chapman 2005b). This method is recommended for datasets with a normal distribution of values, such as those with many observations for each taxon.
- *1.5 x interquartile range (1.5 IQR)* (DIVA-GIS 2005). This method is recommended for datasets with a limited number of observations per taxon (e.g. $n < 20$).

Although these methods are convenient for detecting outliers, they do not guarantee that all detected outliers are, in fact, errors; some outliers may also be valid points. One important detail to consider before deciding to remove the point(s) is the purpose of the study. If the purpose is to model species distribution in order to identify suitable production areas, then using only core records may be preferable and there is no need to include the outliers in the analysis. However, if the purpose of the study is to identify ecotypes that may be adapted to more extreme conditions, atypical points are of interest to the study and it may be useful to keep outliers included in the analysis. As usual, a record of all modifications made to the original file or dataset should be maintained (Chapman 2005a).

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel 	<p>Data Files:</p> <p>Folder 4.2 Quality Control - Atypical points</p> <ul style="list-style-type: none"> • <i>Vcundinamarcensis_outliers</i> (shp, shx, dbf) • <i>Vcandicans_outliers</i> (shp, shx, dbf) • <i>Latin America countries</i> (shp, shx, dbf) <p>For this analysis, you need to have the 2.5 min worldclim climate data imported in DIVA (cf. Section 2.2)</p>

4.2.1. How to identify outliers based on environmental data

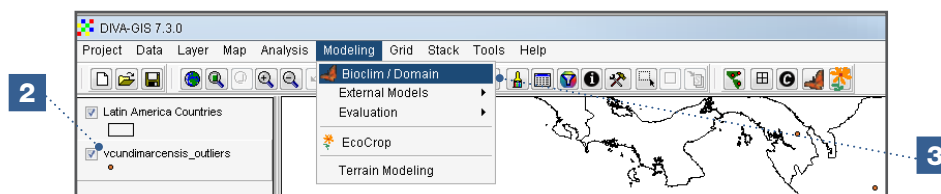
In this analysis you will learn how to identify outliers using DIVA-GIS and data for *V. cundinamarcensis* and *V. candicans* (Scheldeman et al. 2007). Errors were intentionally included in the database for the purpose of the analysis. Errors can be identified where the species name has been replaced with the word 'Error' in the dataset. The *V. cundinamarcensis* dataset is large enough (144 observations) to be revised using the *Reverse jackknife* method, while the dataset for *V. candicans* is smaller (17 points) and will need to be revised using the *1.5 IQR* method.

Steps:

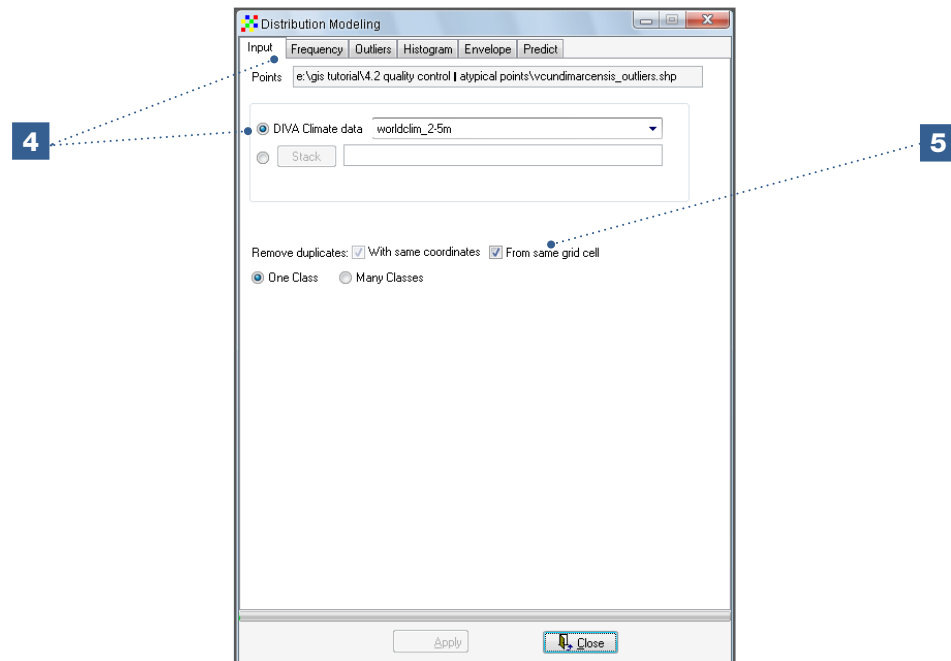
1. Open the vector (*.shp) files, *Vcundinamarcensis_outliers.shp* and *Latin America countries* in DIVA-GIS.



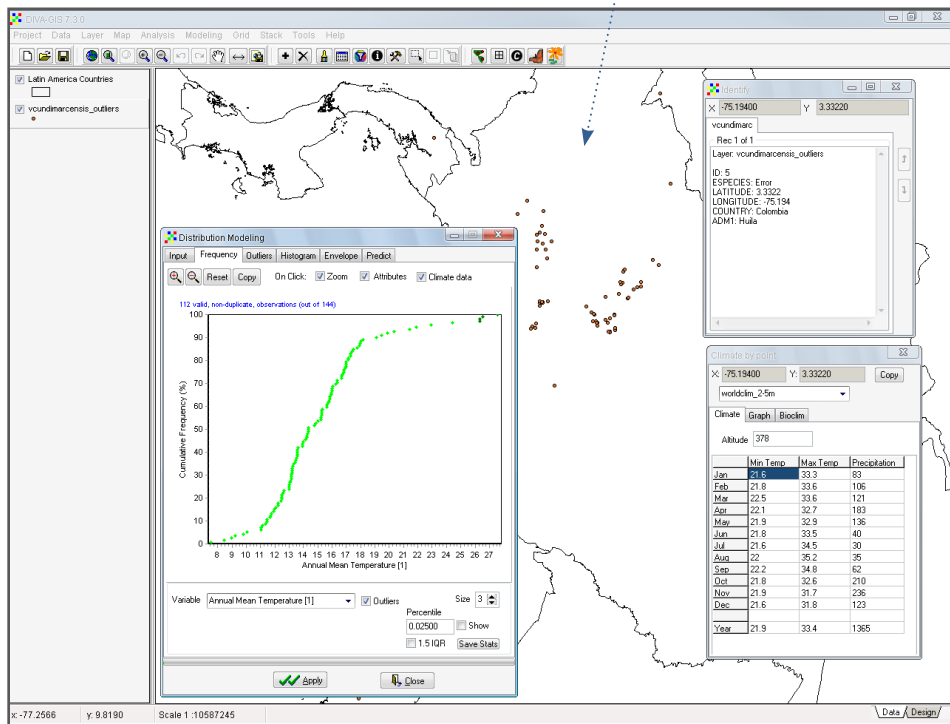
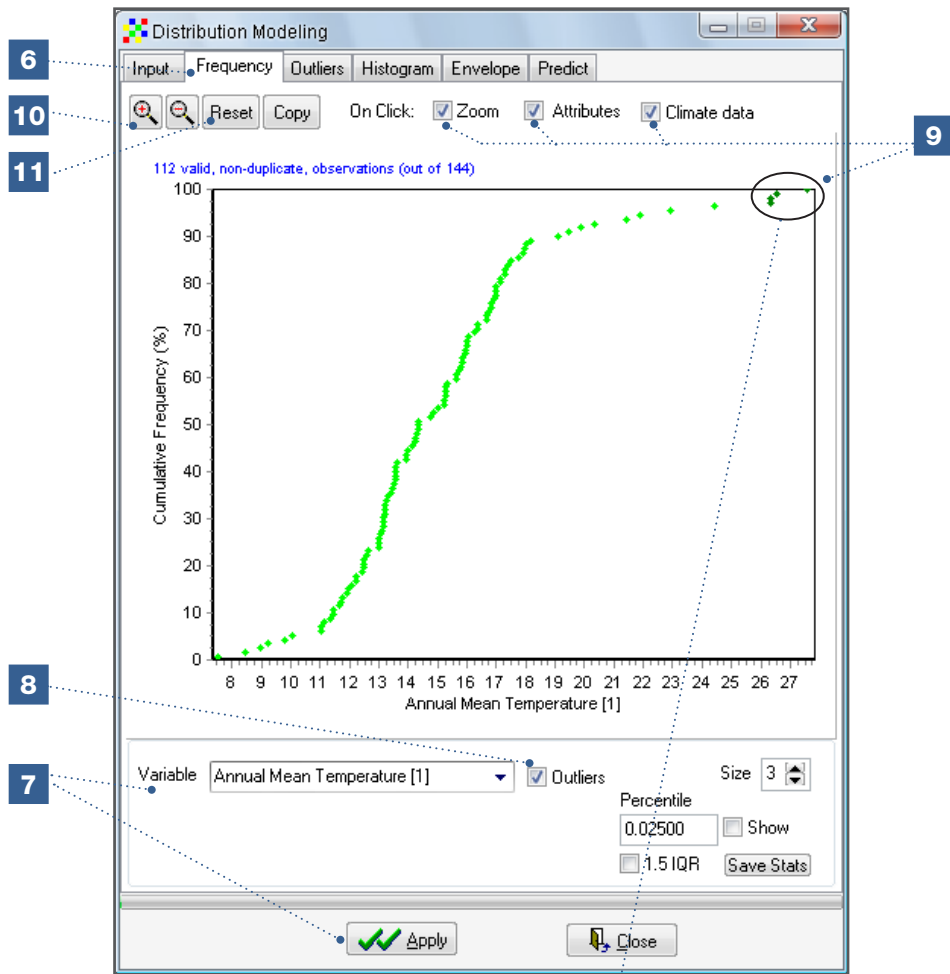
2. To detect atypical points using the *Reverse jackknife* method, select the point file by clicking on the folder in the legend.
3. Go to *Bioclim/Domain* in the *Modeling* menu.



4. The *Vcundinamarcensis_outlier.shp* file and the *worldclim_2-5min* climate data file should appear automatically in the *Points* and *DIVA Climate data* windows. If they do not appear, open them manually.
5. Keep the *From same grid cell* option checked to exclude duplicates in the climatic variables cells.



6. Go to the *Frequency* tab to understand the climate distribution of the presence points of *V. cundinamarcensis*.
7. Select the climatic variable of interest. For this analysis, the selected variable is the *Annual Mean Temperature* file. Click *Apply* to see the results on the graph.
8. Using the *Outliers* option, atypical data according the *Reverse jackknife* method will be highlighted in dark green.
9. Check the *Zoom*, *Attributes* and *Climate data* boxes and click on any point in the graph. The corresponding presence and climate data for that point will be displayed and the point will be highlighted on the map for a few seconds.
10. To adjust the size of the graph, use the *Zoom in* or *Zoom out* buttons, located on the upper left-hand of the menu.
11. The *Copy* option allows you to copy and paste the graph into any other type of document.



Note

Sometimes there is more than one outlier at the same location. You can recognize this in the *Attributes* box when more than one presence point is indicated (e.g. *Rec 1 of 2*). If this is the case, you can view in the *Attributes* box, the passport data of the different outliers present at the same location by clicking on the arrows on the right-hand of the window where the passport data of each point is presented.

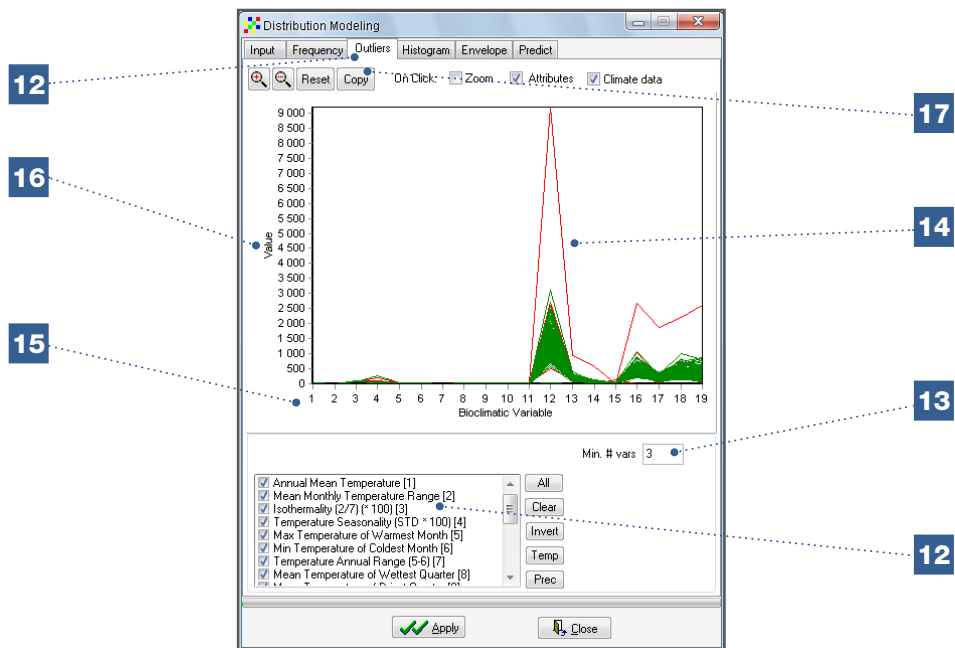
Normally, a point would be classified as atypical after using a combination of various bioclimatic variables in the analysis. Using the *Outliers* tab, you can define the conditions and the set of bioclimatic variables important for the occurrence of a specific species; you can also indicate the minimum environmental variables required in order to consider a presence point as atypical (with extreme values in different bioclimatic variables).

12. Go the *Outliers* tab and select the climatic variables you would like to include in the atypical points analysis. For this analysis, all variables have been selected. You can also select a set of climatic variables you consider key for the occurrence of the species under study.
13. Select the minimum number of variables for which a presence point should have atypical values to be considered an outlier. In this analysis, three variables are selected.
14. The lines on the graph represent presence points. Red lines represent outliers under the conditions specified. Double-click on one of the lines, making sure that the options: (a) *Attributes* and (b) *Climate data* are selected.

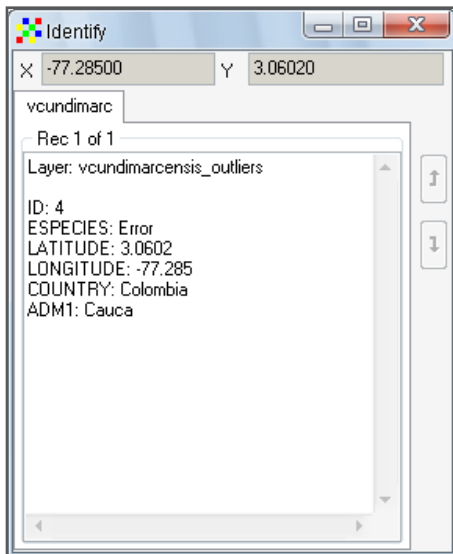
The following information should display automatically when clicking on a line:

- a. Presence data (*Attributes*)
 - b. Climate data.
15. The X-axis of the graph shows the bioclimatic variables: 1) *Mean Annual Temperature*; 2) *Mean Monthly Temperature Range*, etc. (cf. table of Bioclimatic variables in Chapter 2).
 16. The Y-axis of the graph shows the values of each bioclimatic variable for each presence point.

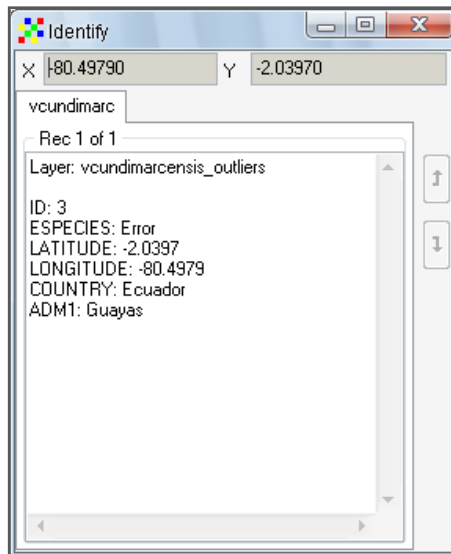
17. The *Copy* option allows you to copy and paste the graph to any other type of document.



14a



14b



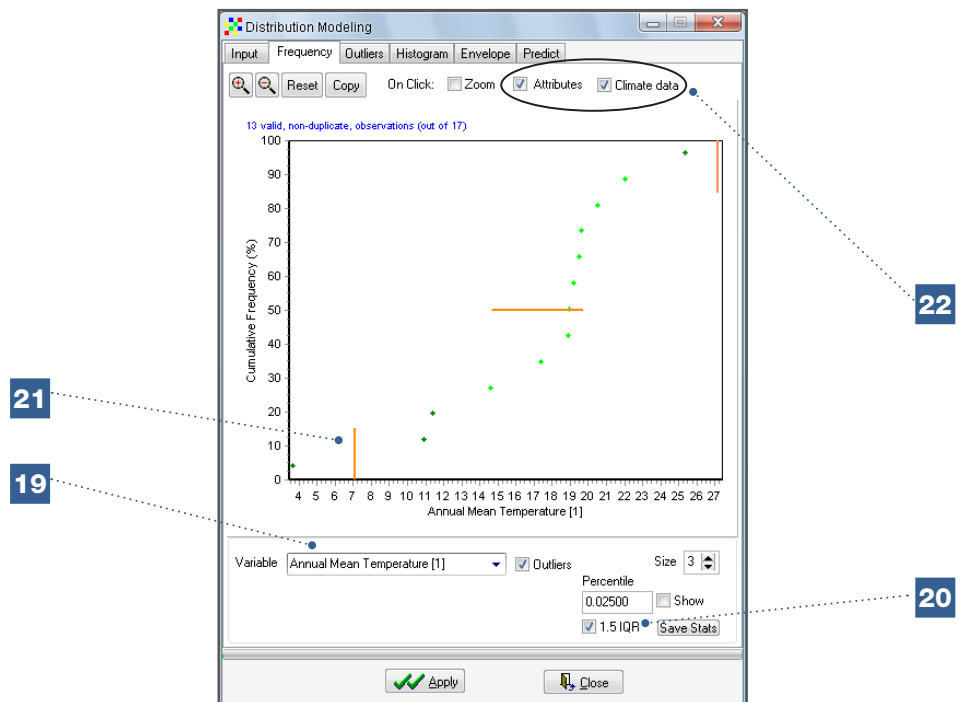
Individual Task: Find all erroneous points in the *Vasconcellea cundinamarcensis* presence point database (the points indicated as errors).

Presence points of small datasets

Presence points of small datasets can be tested to determine if they are atypical by using the *1.5 IQR* method available under the *Frequency* option.

18. To identify atypical points in the *V. candicans* dataset using the *1.5 IQR* method, start by following Steps 1 to 6 above using the new dataset.

19. Define the climate variable based on which you will apply the 1.5 IQR outlier method. In this analysis, select *Annual Mean Temperature*.
20. Select the 1.5 IQR window to establish the limits of the range.
21. Atypical points are those falling outside the 1.5 IQR limits defined by the two outer lines.
22. When the *Climate data* and *Attributes* boxes are checked, the following information should be displayed automatically when clicking on a dot:
 - a. Presence data (*Attributes*)
 - b. Climate data.



22a

22b

	Min Temp	Max Temp	Precipitation
Jan	-1	11	119
Feb	-0.4	10.5	134
Mar	-0.6	10.8	161
Apr	-2	10.7	87
May	-4.8	10.2	26
Jun	-6.8	11.2	6
Jul	-7	11.8	5
Aug	-7.1	11.1	11
Sep	-5.1	11.3	40
Oct	-3.8	10.9	86
Nov	-3	11.4	78
Dec	-2.1	11.9	87
Year	-3.6	11.1	840

Note

Under the *Frequency* tab, you can simultaneously highlight outliers identified with the *Reverse jackknife (Outliers)* method and those identified through the *1.5IQR* method. Note that the *Reverse jackknife* method is not appropriate for smaller datasets as it is too rigorous and would significantly reduce the number of observations (including correct data).

Individual Task: Detect the erroneous points in the presence point data of *Vasconcellea candicans* using different climate variables to apply the 1.5IQR method.

References

- Chapman AD. 2005a. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
- Chapman AD. 2005b. Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
- DIVA-GIS. 2005. User Manual, version 5.2 [on line]. Available from: http://www.diva-gis.org/docs/DIVA-IS5_manual.pdf. Date accessed: October 2010.
- Hijmans RJ, Garrett KA, Huamán Z, Zhang DP, Schreuder M, Bonierbale M. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conservation Biology* 14(6): 1755-1765.
- Scheldeman X, Willemen L, Coppens D'eeckenbrugge G, Romeijn-Peeters E, Restrepo MT, Romero Motoche J, Jimenez D, Lobo M, Medina CI, Reyes C, Rodriguez D, Ocampo JA, Van Damme P, Goetghebeur P. 2007. Distribution, diversity and environmental adaptation of highland papaya (*Vasconcellea* spp.) in tropical and subtropical America. *Biodiversity and Conservation* 16(6): 1867-1884.
- Williams PH, Margules CR, Hilbert, DW. 2002. Data requirements and data sources for biodiversity priority area and selection. *Journal of Bioscience* 27 (4): 327-338.



Section B

Data Analysis

Chapter 5

Spatial analysis of diversity for conservation planning

Statements regarding the current state of biodiversity are frequently made in articles and reports focusing on conservation. These include comments such as '75 % of diversity has been lost' or 'this site is rich in diversity'. However, in order for such claims to be used in the policy- and decision-making process, they must be credible and based on well-established methodologies for assessing levels of diversity and comparable measurements. Although this may sound evident and simple, many questions still exist as to the best ways to measure diversity and how to compare and combine results from different studies. Due to the availability and use of various analytical options, confusion as to the accuracy of results often arises among the multitude of studies; results can be difficult to compare, negatively affecting their validity. This chapter aims to enhance the reader's understanding of the options available when undertaking spatial analysis of diversity and the associated implications for conservation; the chapter provides guidance for selecting appropriate methodologies to conduct analyses and interpret results.

The first challenge when conducting any type of diversity analysis is determining the appropriate level at which to work. Plant biodiversity is studied at the community level (ecosystem), the species level and the genetic level. This manual focuses solely on the study of diversity at the species and genetic levels. At the species level, the observed unit of diversity is the species, measured as present or absent in a certain location. In terms of genetic diversity, the observed unit of diversity may either be a phenotypic trait (the product of gene/genes expression) or a DNA base pair composition (analyzing neutral markers or known functional DNA, based on sequences or molecular weights).

The diversity (of species, varieties, alleles) in distinct subunits within a study area (known as *alpha* diversity) is the principal subject of the spatial analysis of diversity. Subunits of a study area may refer to previously identified sites, administrative units or raster cells of any chosen size. In this manual, raster cells are used to represent subunits of diversity. The advantage of using raster cells is that these allow the comparison of species, trait or allele presence/absence between subunits of similar geographical size throughout the extent of the study area. In some cases vector data is used, for example to compare the number of species between different countries.

The most direct measurement of *alpha* diversity results from counting the number of observed diversity units (e.g. the number of species or the number of alleles per subunit of the study area). Referred to as *richness*, this type of measurement is straightforward and fairly easy to interpret (see Sections 5.1 and 5.3). The analyses in this chapter will focus mainly on this type of measurement as it is widely used to assess diversity at the species level. The richness of alleles is also considered a key measurement for analyzing the conservation of the genetic diversity of a species of interest (Frankel et al. 1995; Petit et al. 1998).

A drawback of the richness measurement, however, is that it depends on the number of samples taken within each subunit of the study area. It is common to find higher levels of diversity in instances where many samples have been collected, while under-sampled areas often appear to have lower levels of diversity; such results are not always accurate

(Hijmans et al. 2000). With the use of more complex analytical methods, this error may be remedied; for example, by using rarefaction, which recalculates the diversity measured at each subunit of the study area to a standardized identical number of samples (Petit et al. 1998; Leberg 2002) (see Section 5.3). Still, the richness of a specific site might be difficult to assess using this alternative, especially when a limited number of observations are available. This is often the case when a high resolution raster (e.g. with cells of 1 km or 5 km) is used for a spatial diversity analysis at a large scale (for example, the analyses described in Sections 5.1, 5.3 and 5.4). In this instance, it is impossible that each of the raster cells (often in the thousands) possess a high number of observations. Chapter 6 discusses how to apply species distribution modelling to address the issue of incomplete sampling or sampling bias.

Another disadvantage of measuring richness is that this methodology does not consider the relative proportions of the number of observed units of diversity. For example, at a site where 150 observations are recorded for a total of three species, there may be 50 observations for each species or 148 observations for one species and only one observation for each of the remaining two. The first situation would clearly be more diverse than the second. Several indices, such as the *Shannon* and *Simpson* indices, have been developed to assess diversity taking into account the respective proportions of each species in the study area, which is also referred to as the measurement of *evenness*. In the case of diversity analysis using molecular markers, indices of allelic frequencies (as a measure of evenness) are used in order to ensure relative proportions of different alleles on each locus in the study area are considered. However, a challenge when utilizing indices which take proportions into account is that these are not appropriate for cells with a limited number of observations, which is generally the case with analyses based on high resolution rasters. Thus, when working with high resolution rasters, richness may still be the most appropriate method to measure diversity.

In addition to richness and evenness, other measurements to assess *alpha* diversity, specific to certain types of data, may be applied. When measuring morphological traits, statistical parameters such as variance and the coefficient of variation can be calculated to determine levels of phenotypic diversity at a given site (see Section 5.2). Further, the application of a multivariate analysis results in distances between individuals in multivariate space (i.e. Euclidian distances). This information can then be used to group similar individuals and undertake subsequent richness analyses. In the case of molecular marker data, in addition to the previously described diversity analysis based on allelic richness, specific genetic parameters (e.g. heterozygosity or the distribution of locally common alleles) for describing differences in intra-specific diversity between subunits can be used as well.

Complementary to *alpha* diversity is *beta* diversity, which focuses on divergence in species, trait or allelic composition between different subunits of the study area. For example, to understand how genetic diversity is spatially structured, subunits of the study area can be clustered based on the genetic similarity of cell composition (see Section 5.3). In Section 5.4 (reserve selection), an analysis is presented on how to combine measures of both *alpha* and *beta* diversity to prioritize sites for conservation. In the process of selecting areas for conservation, emphasis is most often placed on conserving the highest number of species (or alleles; see Petit et al. 1998). It is, however, important to realize that focusing conservation only on those sites with the highest levels of diversity may lead to a failure to identify threatened species found only at sites with generally low levels of diversity (e.g. high mountain ecosystems which reveal a low number of species, but where such species are unique and not found in other ecosystems).

Usually, biodiversity studies assess the status of species or genetic diversity at a specific point in time. While dynamic changes in biodiversity can be detected when site data is collected several times, this topic is beyond the scope of the manual.

5.1. Species richness

Many diversity analyses focus on diversity at the species level. As mentioned, richness is the most straightforward method to evaluate (*alpha*) diversity. This section outlines how to undertake this type of analysis.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel 	<p>Data Files:</p> <p>Folder 5.1 Species Diversity</p> <ul style="list-style-type: none"> • <i>Vasconcellea species</i> (shp, shx, dbf) • <i>Latin America countries</i> (shp, shx, dbf)

5.1.1. How to carry out a spatial analysis of species richness

The analysis below uses data from a diversity study of the *Vasconcellea* genus (Scheldeman et al. 2007). The genus *Vasconcellea* has 21 species, all of which are related to common papaya (*Carica papaya*), and its natural populations are distributed throughout Latin America. Due to an ability to adapt to high altitudes, the species are sometimes known as 'highland papayas'. While some species are grown specifically for their fruit (especially in the Andean region), others are used as a source of genetic material for common papaya breeding programmes [e.g. for specific traits such as tolerance to cold or resistance to the papaya ringspot virus (PRSV-p)]. Further, certain species are widely distributed (*V. cundinamarcensis* can be found from Costa Rica to Bolivia) while others, such as *V. palandensis*, have limited distribution areas and are in danger of extinction. Conservation of the genus is important given its potential for both fruit production and papaya breeding and such conservation efforts will greatly benefit from information on the distribution and diversity of the genus.

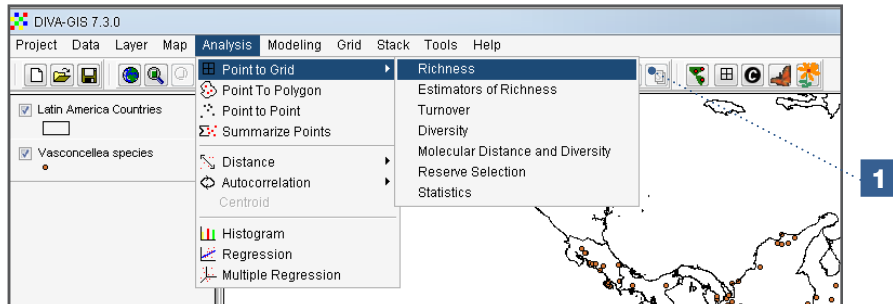
Before starting the analysis, it is important to remember that a sufficient number of observations within a raster cell is necessary in order to undertake reliable diversity analyses. The credibility and accuracy of the final result will depend on the quality of the sampling strategy, although the choice of the cell size used for the analysis (ideally, the cell size should be defined while formulating the sampling strategy) will also influence the quality of the final result. If an analysis is run with cells that are too small, the resulting raster will generate a high resolution map with results of limited value, as each cell will most likely contain too few presence points (often only one) to detect a spatial pattern of species diversity. On the contrary, if raster cells are too large, they will have a sufficient number of observations, but the map will be of poor resolution, complicating its interpretation and use.

The following analysis will use the number of species as the measured unit of diversity. The analysis will be conducted at the regional level (Latin America) and the raster will use a cell of 1 degree x 1 degree (111 km x 111 km at the equator line, see table in Chapter 2). Here, you will learn to use the *Analysis (Point to Grid)* menu in DIVA-GIS to conduct an inter-specific diversity analysis of the *Vasconcellea* genus.

Simple richness analysis in DIVA-GIS based on point to grid analysis

Steps:

1. Start by visualizing two layers: *Vasconcellea data* and *Latin America Countries*. Then, select the layer for the *Vasconcellea* species and go to *Analysis/Point to Grid/Richness*.

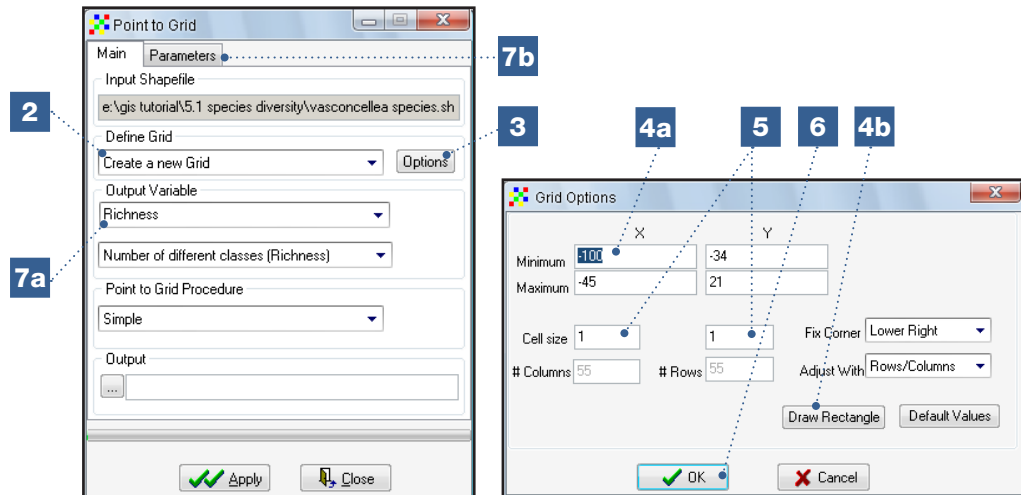


2. Next, define the properties of the raster that will be used for the analysis. The dimensions of the study area as well as the resolution (cell size) must be determined. In the *Point to Grid* window, go to the *Define Grid* option and select *Create a New Grid* (default option).
3. Click on the *Options* window to define the raster properties: origin and extent of the study and the resolution (cell size).
4. The study area can be defined using one of the following options:
 - a. In the *Options* window, manually enter the values for the X-axis and Y-axis (you may wish to select the default options); or
 - b. Draw the extension on the map using the *Draw Rectangle* tool.

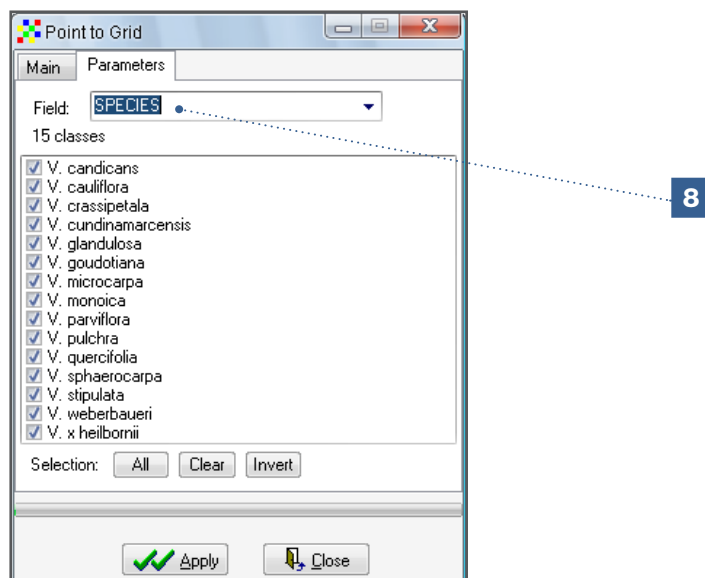
The values used for this analysis are: Min X: -100; Max X: -45 and Min Y: -34; Max Y: 21. These are the default values when the *Vasconcellea* layer is selected at the start of the analysis. Another grid origin will generate a slightly different result.

5. The cell size used in this analysis will be one (1) degree (default value under *Cell Size*), which is equivalent to 111 km at the equator.
6. Click *OK* to accept the values.

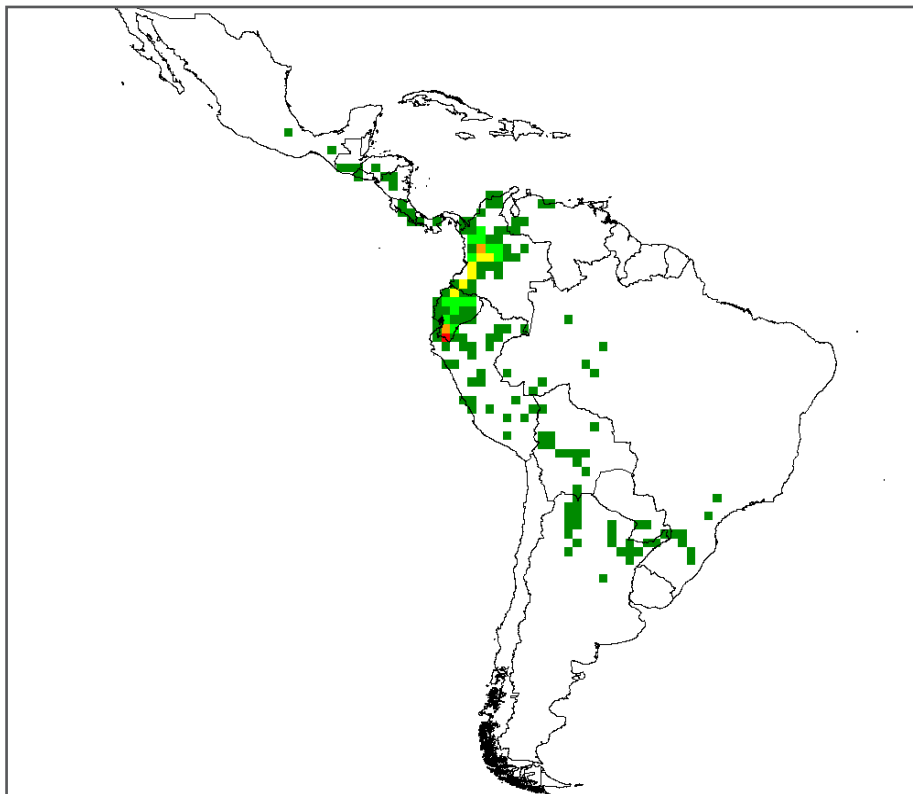
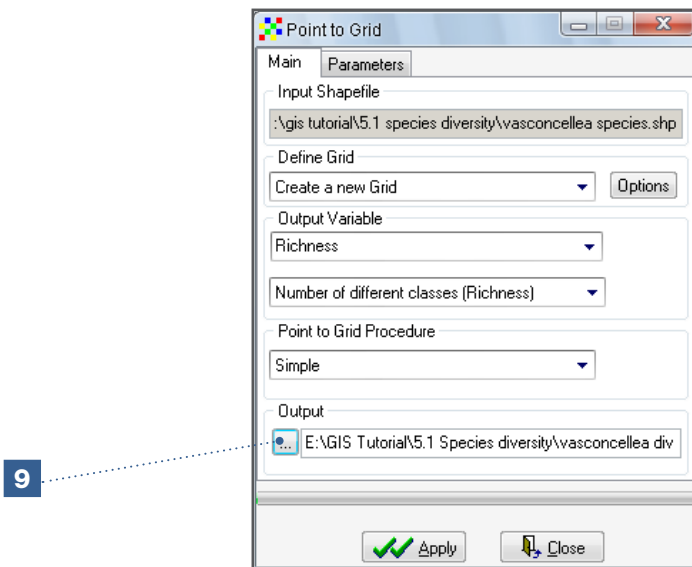
7. Return to the *Point to Grid* window to select the type of analysis to be carried out. Here, we will undertake a species richness analysis.
 - a. First, select the Output Variable: *Number of different classes (Richness)*
 - b. Click on the *Parameters* tab to select the units.



8. In the *Parameters* tab, under the *Field* option, indicate the parameter you wish to analyze. For this analysis, the parameter will be *Species*, in order to analyze the species richness. You are given the option to exclude specific species from the analysis (by un-checking the boxes in front of each species). In this analysis all species will be included.



- Click on the *Main* tab and select the button to the left of the *Output* box (the button with the ellipsis). Enter the name of the file to be saved and its file path. Finally, click on *Apply*. The resulting raster will show the number of species observed in each cell.



Results of the analysis (after moving layers) show that cells in southern Ecuador and central Colombia contain up to nine different *Vasconcellea* species indicating that, overall, Ecuador and Colombia possess a higher diversity of this genus as compared to other Latin American countries. Data used in this analysis correspond to a subset of information used in the study conducted by Scheldeman et al. (2007) and are therefore slightly different from the actual results given in the paper. In the complete study, rare species endemic to southern Ecuador are also included, thus contributing to high levels of diversity in this zone.

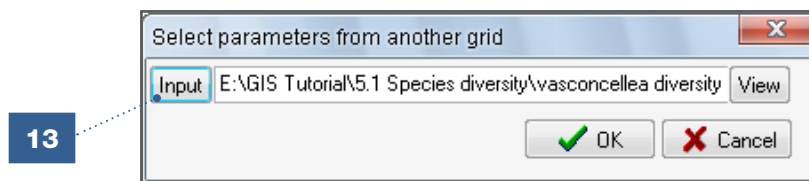
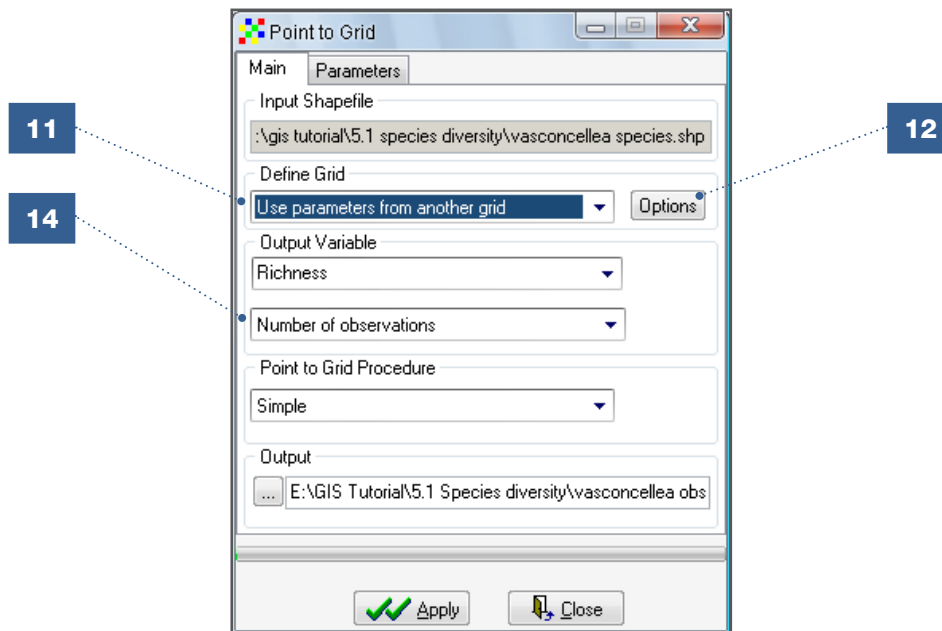
Individual Task: Use DIVA-GIS tools to display a meaningful legend of species richness (see Analysis 3.1.2).

To undertake this task, you will need to know the value of each cell, which can be determined by clicking on the *Information* button or using the arrow, together with the information shown on the status bar of the map (see Chapter 3). The final result should be a map similar to the one below.

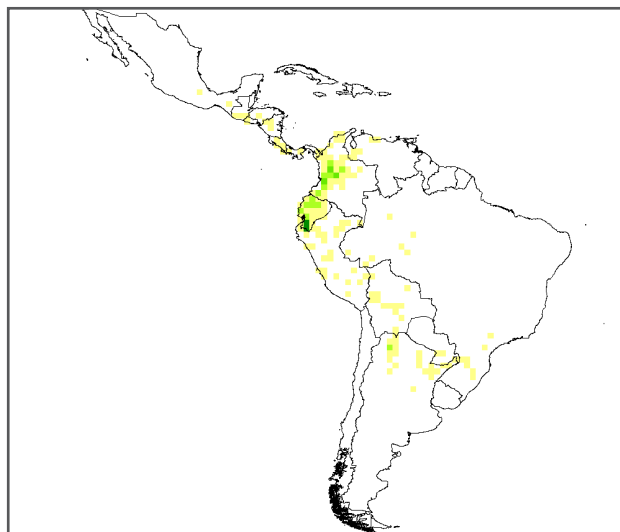


10. Check the number of observations in each cell. To compare results of the previous and current analyses, it is important to use the same raster definitions (see Chapter 3). Go to the *Analysis/Point to Grid/Richness* (see Step 1).
11. Under the *Define Grid* option, select *Use parameters from another grid*.
12. Click on the *Options* button to select the raster from which you wish to use the parameters (the one created in the steps above).
13. Using the *Input* button, select the richness raster file (*.grd) created in the steps above; click on *OK*.

14. Return to the *Point to Grid* window and select *Number of Observations* in the second window of the *Output variable*.



15. Even though this analysis focuses on the number of different observations per cell (which is independent of the observed unit of diversity being analysed), a field must still be indicated in the *Parameters* window. This analysis will use the *Species* field, as was done in Step 8.
16. Indicate an appropriate name for the file (click on the button left to the *Output* box) and click *Apply*.

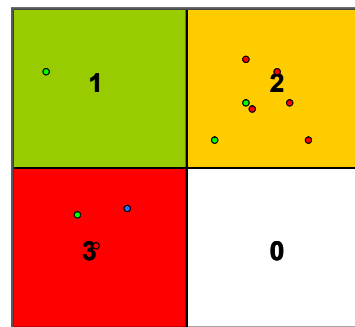


The result (after editing the raster legend as explained in Chapter 3) of this analysis reveals that most observations originate from southern Ecuador. This is a typical situation in which areas known to have high levels of diversity are often preferred sites for botanists and researchers, resulting in more intense sampling efforts and a higher number of observations. Such uneven sampling often takes place in national parks near major cities or in zones with high endemism. This is a common problem and is known as *sampling bias*¹. With the *rarefaction* method, explained in Section 5.3, this problem can be partially addressed, but using this method does result in the loss of certain observations. In Section 6.4, improving diversity studies using species distribution modelling is explained. The *Point to Grid* option in DIVA-GIS also contains several estimators of richness to partially overcome this problem. These tools are also useful to estimate the additional number of species that can occur in each geographic unit of measurement which are not yet observed due to under-sampling. See Section 6.2.2 of the DIVA-GIS Manual, Version 5.2 for more information on these estimators (http://www.diva-gis.org/docs/DIVA-GIS5_manual.pdf). The best solution, though, is to prevent such bias from occurring by ensuring even sampling, to the greatest extent possible.

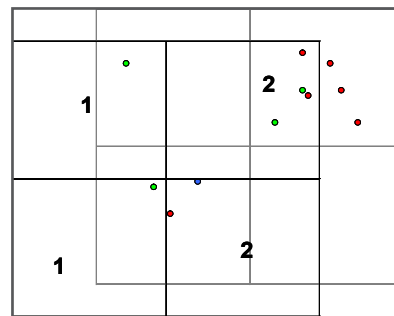
Effects of changing grid origin on the result of a *Point to grid* analysis

A *Point to Grid* Analysis takes into account the observations found in each cell of the raster (e.g. richness looks at the number of observed units of diversity, species for instance, in each cell).

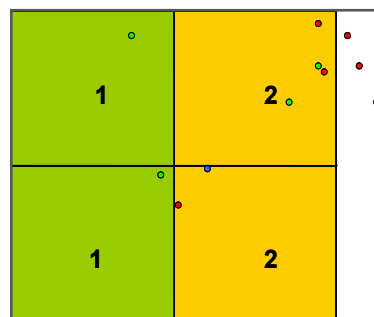
The definition of the raster will obviously influence the result of the analysis. The coming sections explain how differences in raster cell size (resolution) influence the final result. In addition to being defined by the size of its cells, a raster is also defined by its origin (the minimum and maximum X and Y values entered in *Grid Options*). Below is a simple illustration of this effect based on a raster of four cells.



Richness according to raster 1



Use of different raster origin on same data



Richness according to raster 2

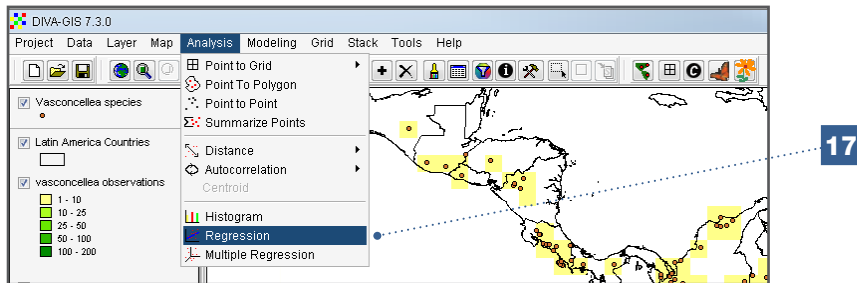
Because of the low number of raster cells, this effect is exaggerated in this example. Results of a more complete sampling will be influenced to a lesser degree by changes in raster properties.

¹ For more information on sampling bias refer to Hijmans et al. (2000).

Visualization of species accumulation curve to assess possible sampling bias

Steps:

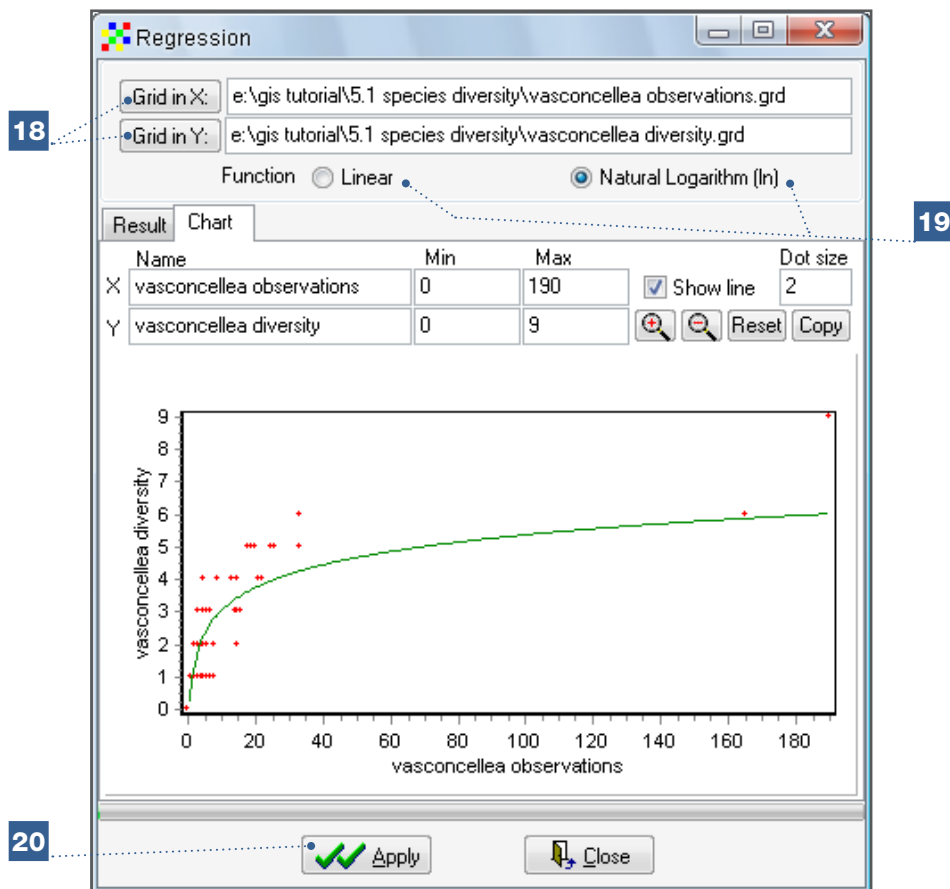
17. A regression allows one to visualize the phenomenon outlined in the previous box. In the *Menu* go to Analysis/Regression.



18. Select the recently-generated layers on species richness and on the number of observations.

19. You can choose between a linear and naturally logarithmic regression. A linear regression is more straightforward, while a logarithmic regression is mathematically complex but may better represent a sample bias as it accounts for the typical 'levelling off' of species accumulation curves as the sampling effort (number of observations) increases.

20. Click *Apply*.



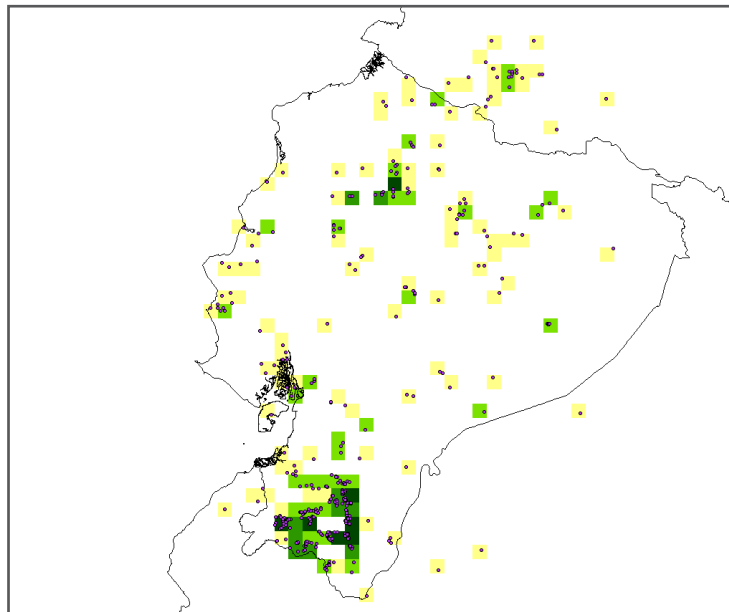
The graph in the *Chart* tab confirms that there is a relationship between the number of samples taken and the number of species observed (with a higher number of sample points more likely to result in a higher richness). As the number of number of observations (sample points) increases, the graph begins to level off, raising questions both about the applied sampling strategy and the cell size used in the analysis (using larger cells and including more observations would be more appropriate).

Only in the instance where the user is familiar with the data and certain it originates from an intense and relatively homogeneous sampling, could the results obtained above be considered accurate and therefore useful. The results must be interpreted carefully if sampling was only partially conducted or fragmented or if the origin of the data is unknown. In these cases, the results may reflect an invalid, biased representation of reality. As mentioned in the introduction of Chapter 4, many spatial biodiversity analyses are partially or totally based on data compiled from herbaria and genebanks, which often reflect non-systematic and uneven sampling (Hijmans et al. 2000; Chapman 2005).

Individual Task: Carefully analyze diversity in Ecuador based on a 10-minute raster (cells of approximately 18 km) (only in Ecuador, using the option Draw Rectangle).
Hint: What will be the desired raster cell size?

Note

Your results may be slightly different from the following as they will depend on the origin of the raster. Since you are only analyzing results for Ecuador, disregard the warning message indicating that there are points outside the selected raster..



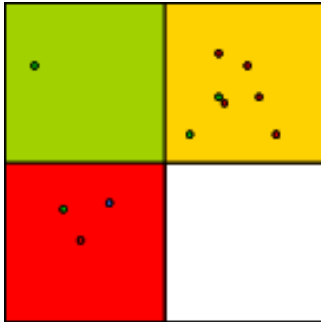
Using the circular neighborhood option for richness analysis

Thus far, the results of analyses have been heavily dependent on the definition of the raster, especially on the size of the cell or resolution. Small cell size generates a higher resolution (detail) but risks losing spatial patterns (for example, when cells are so small that each cell only contains one observation).

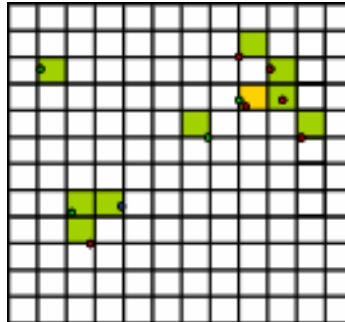
This situation can be improved by applying the *Circular Neighborhood* option, which considers the diversity in adjacent areas. With this option, each cell receives the value of diversity found within a circle with a specified diameter centred on the cell, instead of the value of diversity found within the cell alone.

The concept of circular neighbourhood

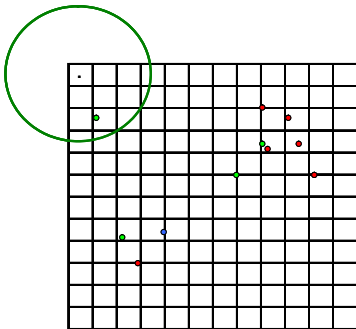
The following diagrams represent the concept of circular neighbourhood, where the circles of different colours indicate different species richness: green indicates a richness value of one, yellow of two and red of three species. When using this method, the results around the margins of the point distribution can be distorted by assigning lower diversity at borders in the case of incomplete sampling (e.g. at country borders).



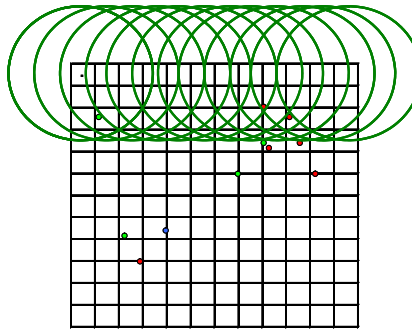
Richness in a one (1) degree cell
Higher diversity is clearly observed in the lower left cell; however, the raster has low resolution.



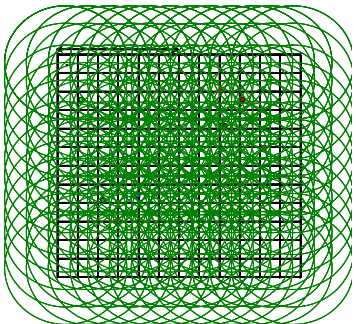
Richness of 10-minute cells
In this case, the raster has more detail (high resolution), but the pattern of diversity has been lost.



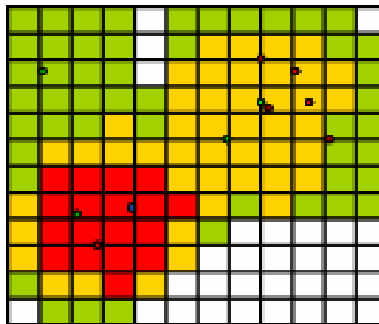
The circular neighbourhood consists of using a fixed-diameter circle centred on each raster cell to assign a value to the cell; in the case of the top left-hand corner cell, the value is one (1), since there is only one species inside the circle shown.



The circular neighbourhood repeats itself for each cell. In this example, the circle has a diameter of one (1) degree, the same size as the cell in the first example.

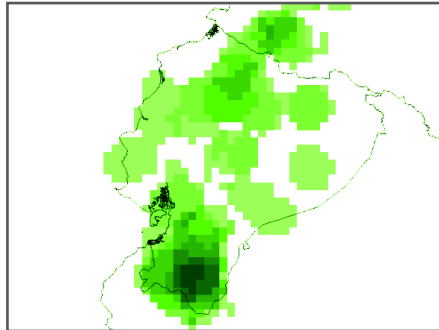


As shown, each observation contributes to the value of a number of different cells.

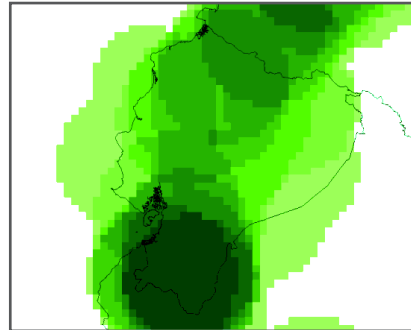


Richness with 10-minute cell and circular neighbourhood of one (1) degree. This methodology allows one to maintain a high degree of resolution without losing the pattern of diversity.

It can be challenging to define the optimal radius of the circle (the best radius is often obtained by trial and error). The example below illustrates that when a circle becomes too large (on the right), the results of the analysis are of little use, having lost the level of detail needed to interpret the output.



Point to Grid Richness Analysis with a 10-minute cell size and circular neighbourhood of one (1) degree.

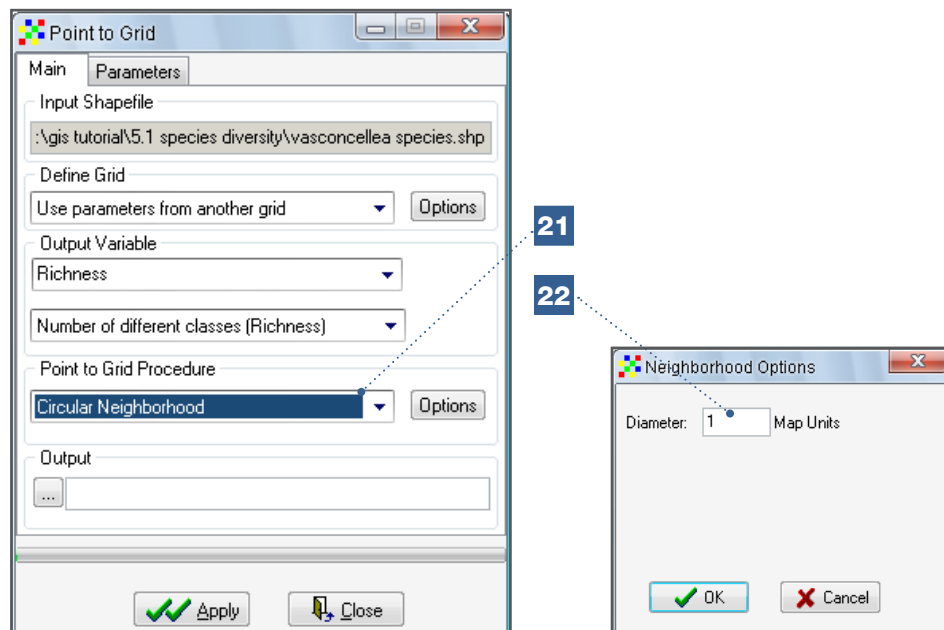


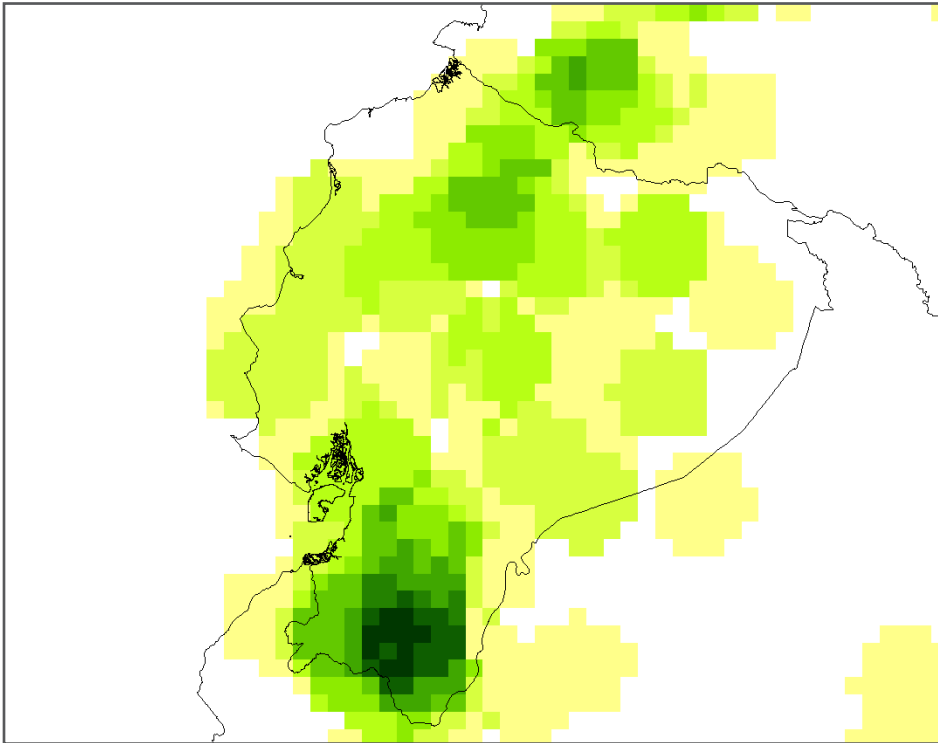
Point to Grid Richness Analysis with a 10-minute cell size and circular neighbourhood of three (3) degrees.

To observe the effect of the *Circular Neighborhood* option and the changes generated in the previous result, the richness analysis will now be repeated using 10-minute cells with a circular neighbourhood of one (1) degree.

Steps:

21. In the *Point to Grid Procedure* window, select the *Circular Neighborhood* option.
22. Use the parameters from the raster generated during the previous analysis (raster cell size 0.1666). Under *Circular Neighborhood Options* enter one (1) as the *Map Unit* to indicate the desired size [in the case of maps based on latitude/longitude coordinates, the map unit is one (1) degree]. Repeat the steps from the previous analysis.





Using the *Circular Neighborhood* option gives you more precise patterns for the existing diversity, along with a relatively high resolution.

Individual Task: Observe what happens if you apply a larger circular neighbourhood (e.g. 5 Map Units).

5.2. Intra-specific diversity analysis based on phenotypic data

Genetic diversity studies, including the analysis of spatial patterns in genetic diversity, are frequently based on molecular marker data (see Section 5.3). However, phenotypic data and, particularly morphological data, can be another indirect source of genetic diversity information. Phenotypic data from a single individual varies as a function either of the genotype (G), the environment (E) or a combination of both (the GxE effect). Some traits, like flower colour, are not influenced by the environment. When using data based on *in situ* characterization, conducted at environmentally heterogeneous locations, it is recommended to focus on these traits. When working on other traits, in order to minimize variation determined by environmental influences, one option is to conduct the characterization outside the original collection site and under controlled, uniform environmental conditions (*ex situ* characterization), be it in the same geographic location (e.g. experimental fields) or in a controlled environment (e.g. greenhouse). Although environmental effects may still play a role in *ex situ* characterization, especially in experimental fields - for example when accessions carry traits which make them better-adapted to the chosen site's conditions - characterization in experimental fields will be more relevant for comparison than those obtained from *in situ* characterization.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Statistical program (optional) 	<p>Data Files:</p> <p>Folder 5.2 Diversity - Phenotypic data</p> <ul style="list-style-type: none"> • <i>Manihot ex situ</i> (shp, shx, dbf) (Cassava <i>ex situ</i> characterization data in Ucuyali) • <i>PER_Adm0; PER_Adm1; PER_Adm2</i> (shp, shx, dbf) (Administrative data for Peru) • <i>Peru_towns</i> (shp, shx, dbf) (Municipalities in Peru) • <i>PER_roads</i> (shp, shx, dbf) (Data on roads in Peru) • <i>PER_water_areas_dcw; PER_water_lines_dcw</i> (shp, shx, dbf) (Data on rivers and bodies of water)

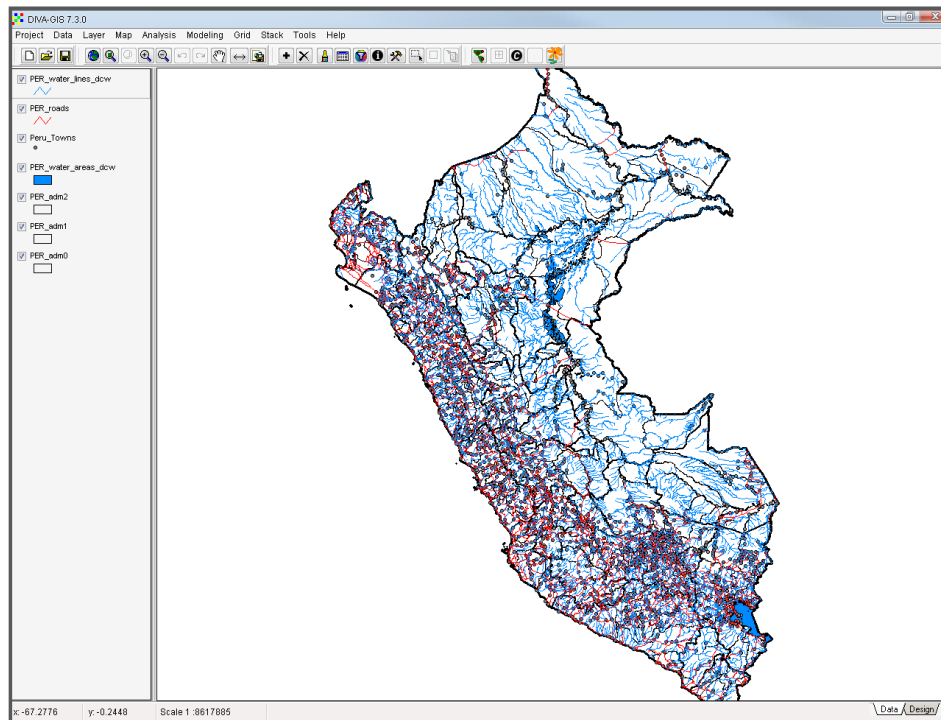
5.2.1. How to carry out a spatial diversity analysis using phenotypic data

The following analysis outlines how to carry out a spatial analysis based on phenotypic data resulting from *ex situ* (in the same experimental field) morphological characterization, combined with passport data (information about the site where material was originally collected). The data used in this analysis comes from a study on the diversity and genetic erosion of cassava (*Manihot esculenta*) in the Peruvian Amazon (Ucayali Region) (Willemens et al. 2007). In this section, you will learn to use the *Analysis (Point to Grid)* menu in DIVA-GIS to undertake an intra-specific diversity analysis based on morphological data.

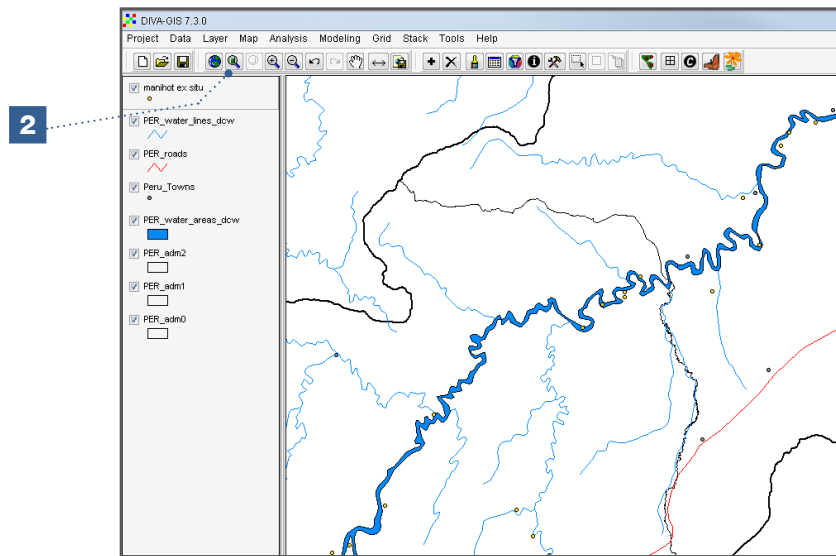
Spatial analysis of statistics of phenotypic data based on point to grid analysis

Steps:

1. Start by reviewing the DIVA-GIS options outlined in Chapter 3 (Basic Elements). Create a map of Peru which shows three administrative levels: country (line width = 3), regions (line width = 2) and districts (line width = 1). Add layers with rivers and bodies of water (in blue), roads (in red) and towns and villages (in gray). The resulting map should look like the following:

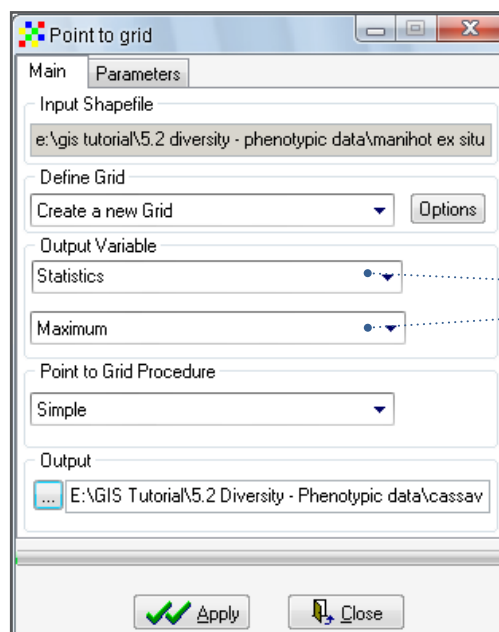


- Now, add the layer with the characterization data of cassava (*Manihot ex situ.shp*) and zoom in on this data using the *Zoom to theme* option.

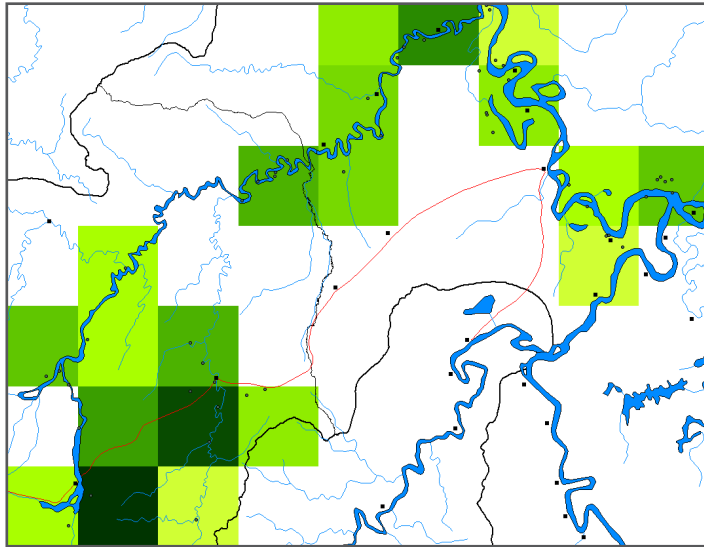


- Explore the *ex situ* characterization values and identify those areas where cassava roots with the heaviest weight were found. The parameter for average fresh root weight (FRW) will be used in this exercise. Select the cassava characterization layer and go to *Analysis/Point to Grid/Statistics*.

Go to *Define Grid File Options/Cell Size* and define a 10-minute raster (cell size = 0.1666). This raster will be used for the remaining analyses in this section. In the *Output Variable* window, select the *Maximum* option. In the *Parameters* window select the parameter to be analyzed (in this case FRW). To finish, assign a name to the generated file. In the map, amend the legend for improved visualization and interpretation.

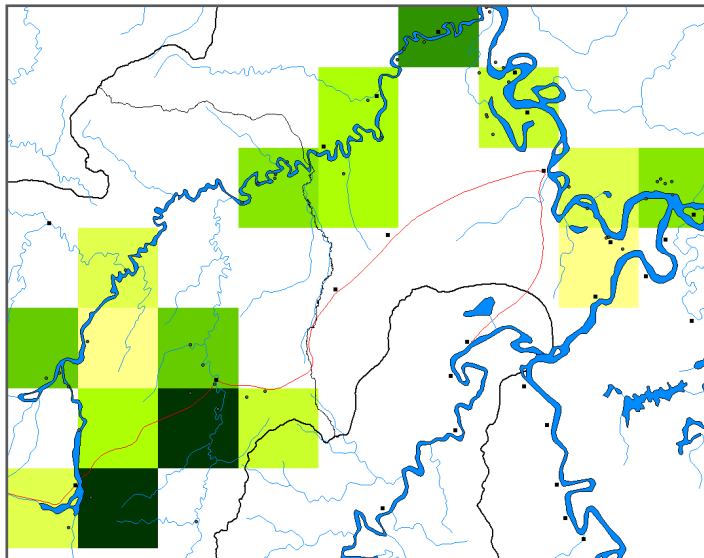


The result (after changing the legend) shows the location of zones of individuals with the highest root weight².



Cassava accessions with the heaviest roots are found mainly in the south-eastern part of the study area. This analysis provides information on a specific characteristic (in this case, fresh root weight) but does not offer information on the areas' diversity. In order to gather this information you will need to look at the range of the analyzed parameter (FRW).

4. Under *Analysis/Point to Grid/Statistics* select the *Range* parameter. Use the same raster as in the previous analysis (FRW). Note that the previously used/established legend can no longer be applied as the calculated parameter (range instead of maximum) has been changed and dimensions will be different. Adjust the legend to the new results.

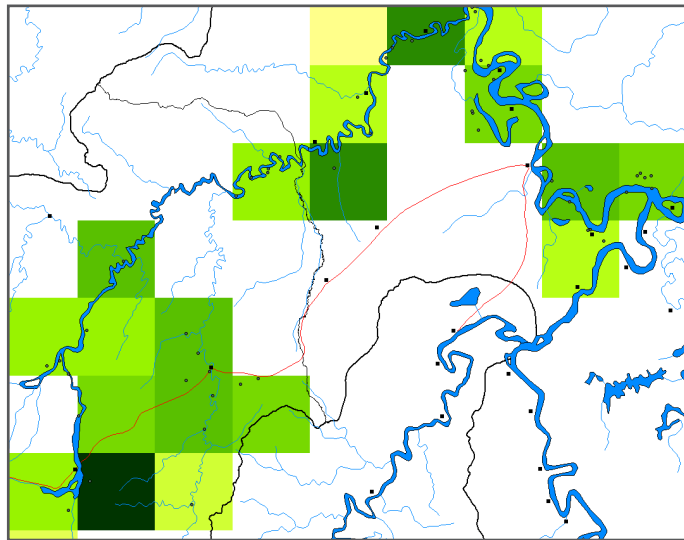


- 2 Results may be slightly different from the map shown, as they will depend on the origin of the raster. This analysis used a raster with the following characteristics: Min X: -78; Max X: -71 / Min Y: -12; Max Y: -5.

The resulting map reveals that those cells corresponding to sites where the heaviest weight is found also correspond to the cells with the greatest ranges in weight. Even though the range provides an idea of the variability within a parameter, it has two disadvantages: firstly, the range only takes into account the extreme values (not the distribution of values); and secondly, the value of the final result will depend on the units of measurement (in this case grams). Combining these results with the results of other trait analyses, such as length of leaf petiole or number of leaf lobes, may be difficult, but is nonetheless important to understand the overall diversity (which needs to include different traits).

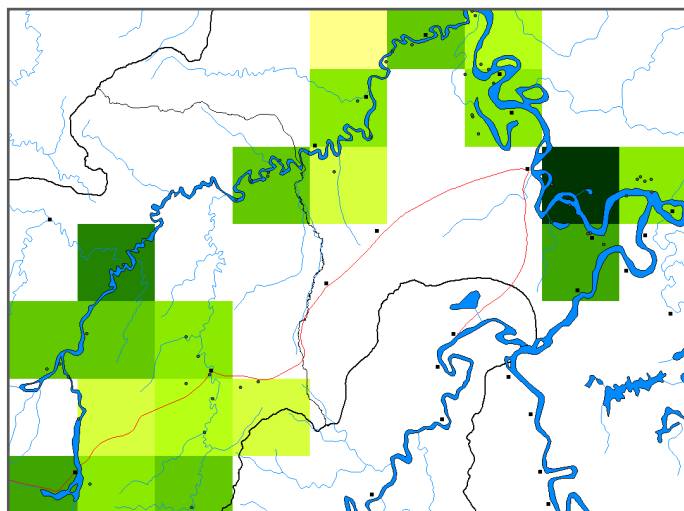
By using a parameter without dimensions [i.e. the coefficient of variation (CV)], which instead considers the distribution of the parameters around the average, such analysis combinations are more feasible.

5. Under *Analysis/Point to Grid/Statistics* select the *Coefficient of Variation* parameter.



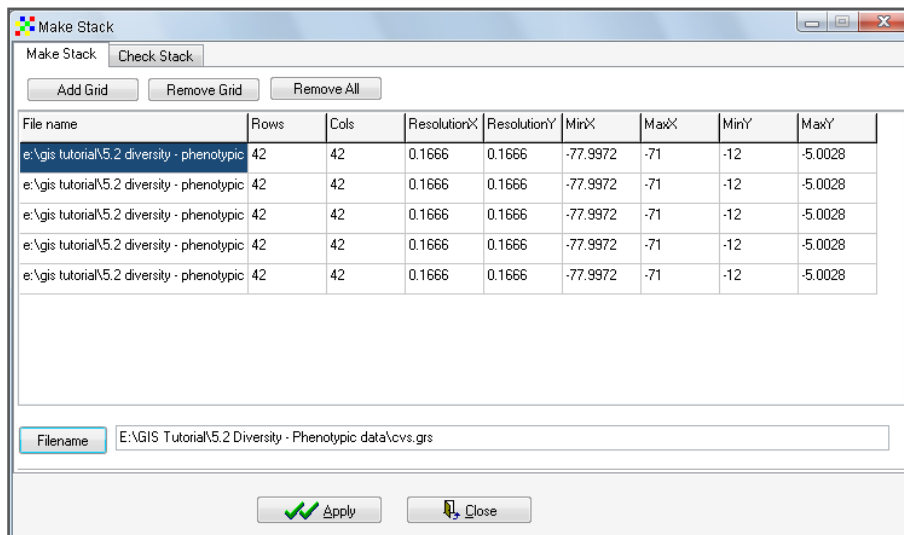
This map provides a better picture of the diversity of the FRW parameter in the different cells.

6. Analyze the CV of another trait, i.e. the length of leaf petiole (LLP). Use the same raster properties from the previous analysis in order to compare the results of the two analyses.

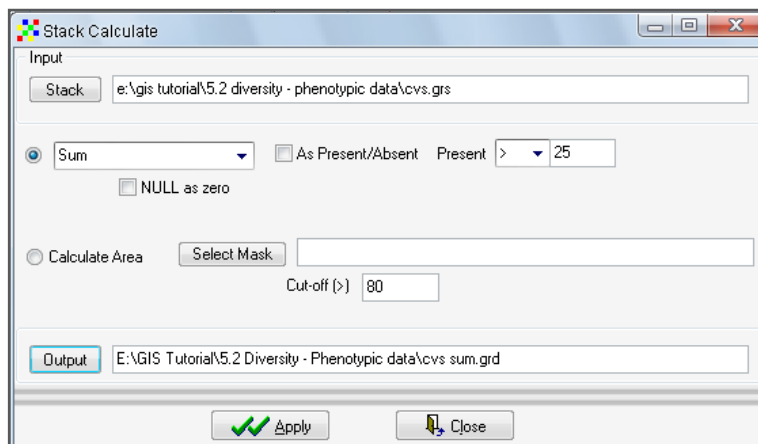


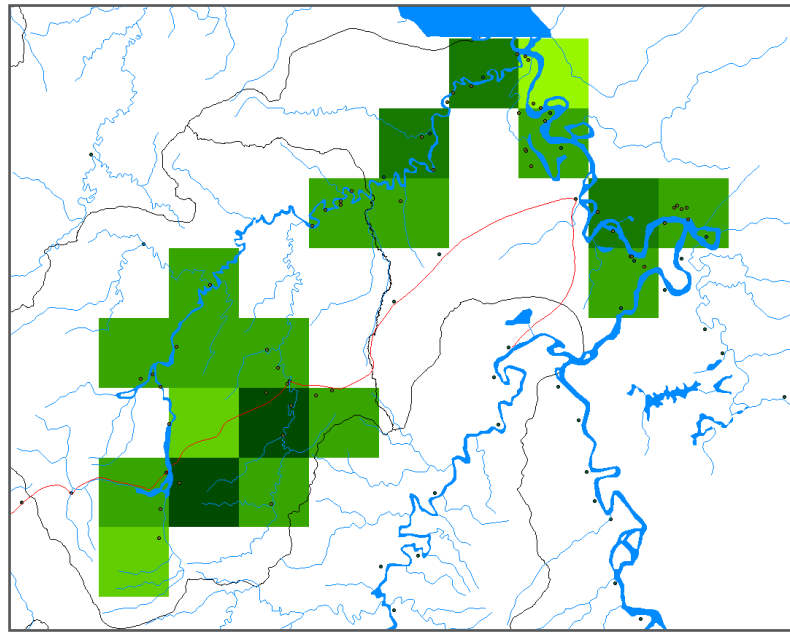
Areas with a high variability for the LLP parameter are different from those for fresh root weight. How do you then define where the sites with the highest cassava diversity are located? The solution lies in combining the different traits. As previously mentioned, unlike the range, the use of the *Coefficient of Variation* enables the combination of different parameters, which is necessary to undertake such an analysis.

7. Five traits will now be included for comparison: the two previously analysed traits, fresh root weight (FRW) and leaf petiole length (LLP); and three new parameters: number of leaf lobes (NLL), distance between nodes (DBN) and days to harvest (DH). Three additional CV layers must be created.
8. The layers generated must now be combined to form one layer by calculating the average CV of the five layers. Go to *Stack/Make Stack* and add the five rasters with the calculated CVs, using the *Add Grid* button. Assign a name to the stack and click on *Apply*. Combining different rasters into a stack is only possible if all rasters have the same properties, as outlined in Chapter 3.



9. Go to *Stack/Calculate* and calculate the mean of the five rasters in the stack.



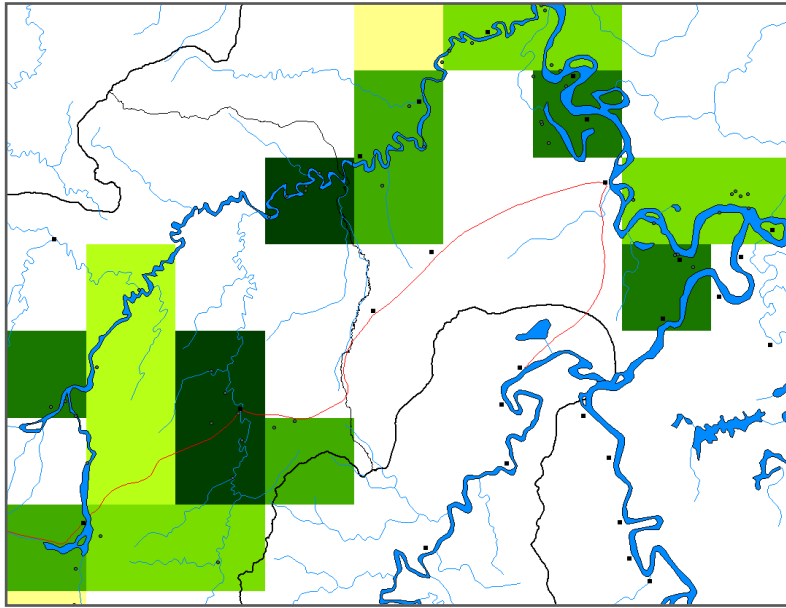


The combination of the layers for the five morphological characters results in a stronger picture of the existing diversity.

Obviously, this is a rather simplified diversity analysis. Ideally, a higher number of quantitative and qualitative traits should be taken into account. To work with more variables, especially qualitative variables, multivariate statistics are required, which are based on additional statistical software. As this is beyond the scope of the manual, please refer to Grum et al. (2007), Kindt et al. (2005) and Mathur et al. (2008) for further information.

Data on phenotypic traits and/or allele presence (reflecting the phenotype and genotype respectively of an individual) can be used as indicators for diversity at the different subunits of the study area, as well as to compare, directly (alleles) or indirectly (traits), genetic similarities between individuals. These similarities can be used in multivariate analyses, such as a cluster analysis. To illustrate this practice, the *manihot ex situ.shp* file includes the results of a cluster analysis. The variable, *Cluster*, shows the cluster to which each individual belongs, based on similarities in traits. This information enables one to conduct a diversity analysis like the one illustrated in Section 5.1, but using the number of clusters as a unit of diversity, instead of different species. In this case, diversity is defined by the number of different clusters found in a site.

Individual Task: Conduct a simple richness analysis based on the results of the cluster analysis (variable: *Cluster*), with the same raster used in previous analyses.



There will be differences between the results of analyses based on the coefficient of variation and those resulting from multivariate analyses. The cluster analysis uses a more complex multivariate statistical methodology and a greater number of variables (Willemens et al. 2007). Nonetheless, the diversity tendencies revealed in the results of the two types of analyses are the same: low diversity exists around the city of Pucallpa (bordering the river where the road to Lima begins) and a very high level of diversity is present in the central zone close to the road to Lima (lower left corner of the map). The cell with the highest diversity is the same in both analyses. As discussed in Section 5.1, the circular neighbourhood technique may also be used for such analyses.

5.3. Intra-specific diversity analysis based on molecular marker data

As noted above, phenotypic/morphological data can be used to conduct intra-specific diversity analyses; however, the influence of the environment where the characterization was conducted will always affect the results. Allelic composition, or gene sequences of plant individuals, are not influenced by such environmental factors (a change in environmental conditions, e.g. wet vs. dry year, does not alter DNA base pair composition but will alter phenotypic appearance, e.g. leaf size or growth). Therefore, molecular markers are the measurements of choice to carry out an analysis of intra-specific diversity.

Although molecular markers are, for the most part, not directly related to functional genes, it can be anticipated that a high diversity based on molecular markers also indicates a high abundance of useful genes. The following section outlines a basic spatial analysis of molecular marker data, whereby microsatellites (SSRs), a widely used co-dominant marker, are utilized. It should be noted, though, that any type of molecular marker data, e.g. AFLPs, can be used to conduct a diversity analysis, provided a unique identity can be given to each variation in the DNA composition.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION

Programmes:

- DIVA-GIS
- Excel

Data Files:

- Folder 5.3 Diversity - Molecular marker data
- *SSR cherimoya rand column (shp, shx, dbf)*
 - *Latin America Countries (shp, shx, dbf)*

5.3.1. How to carry out a spatial diversity analysis using molecular marker data

The data used in the following analysis is from the CHERLA project, which included a major study on the diversity of cherimoya in its centre of origin (for more information see: www.cherla.com). Given that the final results of this study have not yet been published, only a subset of the data will be used, with coordinates randomly modified to recreate a hypothetical scenario.

Principles of spatial diversity analysis using molecular marker data are very similar to those of an analysis at the species level. In this analysis we use microsatellite (SSR) marker data. SSRs are short tandem repeats of base pairs, highly variable and evenly distributed throughout the genome (Hajeer et al. 2000; De Vicente et al. 2004a). SSR analysis looks at differences in length of microsatellite regions (usually not associated to functional genes, i.e. neutral). These differences in length, further referred to as different alleles, are the observed units of diversity in this analysis.

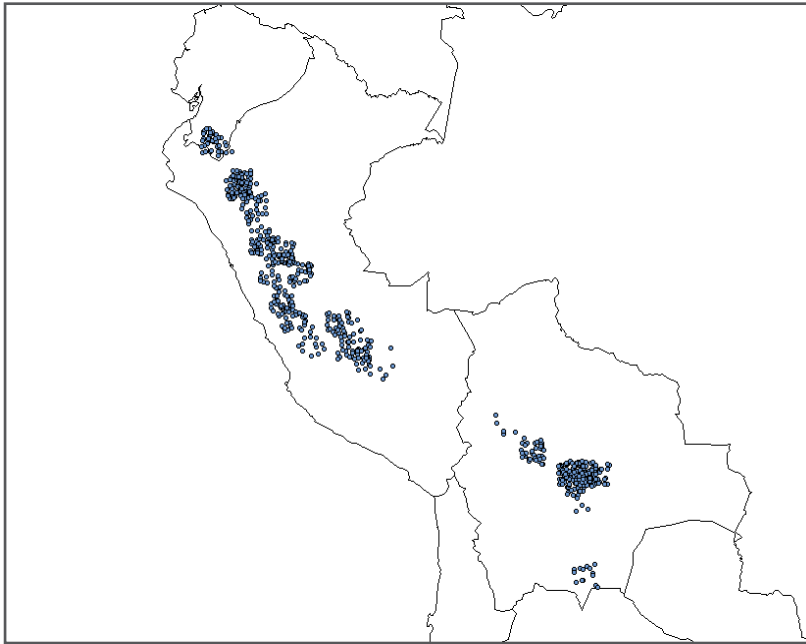
To carry out a spatial analysis, the molecular marker data must be formatted in such a way that each allele includes georeferenced information. In the following analysis, each allele is formatted according to the following: microsatellite code + weight of base pairs (e.g. SSR1-293). Review the example *SSRs cherimoya rand column.dbf* in Excel to become familiar with the table format.

ID	X	Y	SSRS
2	309	-65.2001	-17.1218 SSR1-293
3	309	-65.2001	-17.1218 SSR1-293
4	309	-65.2001	-17.1218 SSR2-122
5	309	-65.2001	-17.1218 SSR2-128
6	309	-65.2001	-17.1218 SSR3-216
7	309	-65.2001	-17.1218 SSR3-218
8	309	-65.2001	-17.1218 SSR4-154
9	309	-65.2001	-17.1218 SSR4-154
10	309	-65.2001	-17.1218 SSR5-156
11	309	-65.2001	-17.1218 SSR5-156
12	309	-65.2001	-17.1218 SSR6-148
13	309	-65.2001	-17.1218 SSR6-148
14	309	-65.2001	-17.1218 SSR7-183
15	309	-65.2001	-17.1218 SSR7-183
16	309	-65.2001	-17.1218 SSR8-305
17	309	-65.2001	-17.1218 SSR8-316
18	310	-65.1752	-17.1442 SSR1-293
19	310	-65.1752	-17.1442 SSR1-293
20	310	-65.1752	-17.1442 SSR2-122
21	310	-65.1752	-17.1442 SSR2-122
22	310	-65.1752	-17.1442 SSR3-222
23	310	-65.1752	-17.1442 SSR3-230
24	310	-65.1752	-17.1442 SSR4-152
25	310	-65.1752	-17.1442 SSR4-154
26	310	-65.1752	-17.1442 SSR5-156
27	310	-65.1752	-17.1442 SSR5-156
28	310	-65.1752	-17.1442 SSR6-142
29	310	-65.1752	-17.1442 SSR6-142
30	310	-65.1752	-17.1442 SSR7-183
31	310	-65.1752	-17.1442 SSR7-183
32	310	-65.1752	-17.1442 SSR8-305
33	310	-65.1752	-17.1442 SSR8-316
34	311	-64.3412	-17.4599 SSR1-293
35	311	-64.3412	-17.4599 SSR1-293

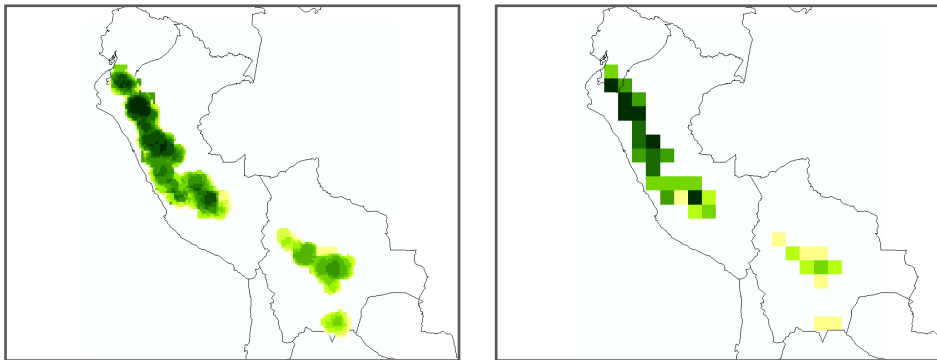
In this section you will learn to:

- Use the *Analysis (Point to Grid and Distance)* option in DIVA-GIS
- Use the *Cluster* option in DIVA-GIS.

Individual Task: Visualize cherimoya sampling sites included in the diversity analysis (see Chapter 3 - Basic Elements).



Individual Task: Conduct an analysis of allelic richness, first using one (1) degree cells, and next based on 10-minute cells with a circular neighbourhood of one (1) degree (see Section 5.1³). Analyze where the highest level of diversity is found.



The rasters generated, shown above, indicate that northern Peru is the region with the highest level of diversity (highest number of different alleles found), while Bolivia maintains the lowest level.

Rarefaction method

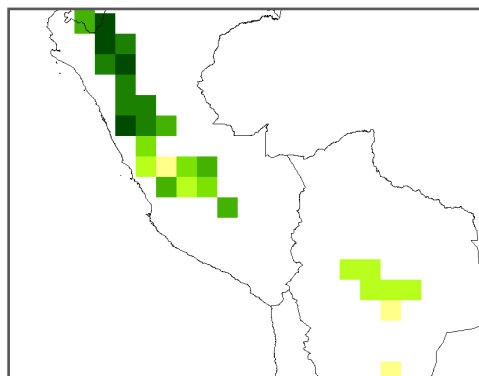
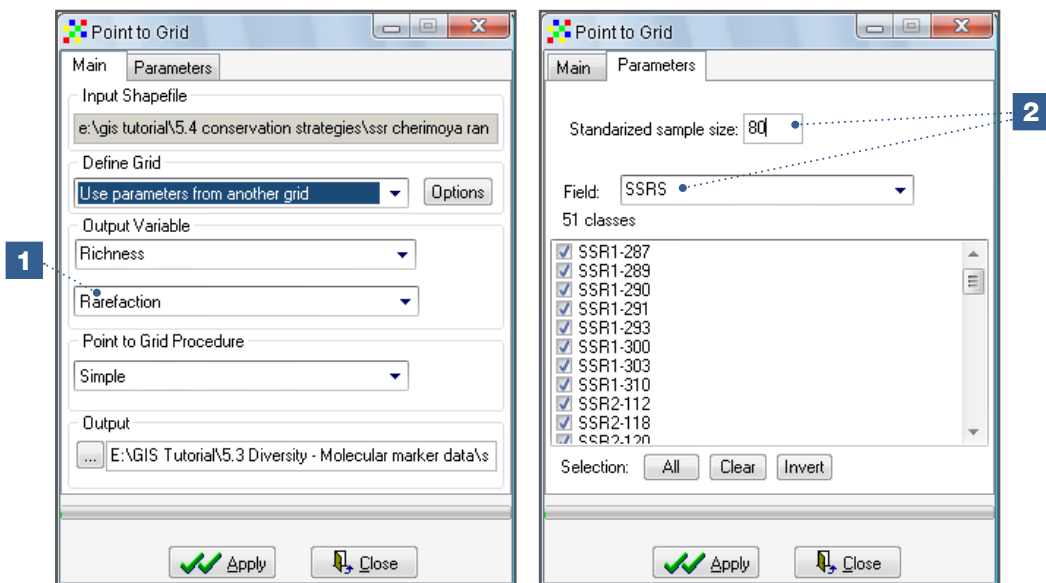
As seen in the first analysis (Analysis 5.1.1) uneven distribution of observations can have a significant impact on the richness analysis. The rarefaction method has been developed to compare richness among cells that have a dissimilar number of observations or samples [for more information, refer to Petit et al. (1998) and Leberg (2002)]. The rarefaction

³ Results may be slightly different from the map shown, as they will depend on the origin of the raster. In this analysis the raster used had the following characteristics: X: Min: -82; Max: -62 / Y: Min: -24; Max: 0.

method recalculates the richness measured in the different cells as if a standard number of observations were made in each cell. Only cells with an equal or higher number of observations than the standardized number are included in the analysis; cells with fewer observations are excluded. The choice of the standardized number of observation is a trade-off between the number of cells included in the analysis and the maximum richness to be calculated. For example, if a small number of observations is chosen, almost all cells will be included in the analysis; however, the maximum diversity a cell will be low as it cannot be higher than the defined number of samples. If a high number of observations is chosen, the number of cells with at least this number of observations (essential to be included in the rarefaction calculation) will be low, resulting in a limited number of cells with a calculated value. The following analysis is based on eight microsatellites, whereby 80 allele observations were used as the fixed sample size (equivalent to 10 homozygous individuals or 5 heterozygous individuals).

Steps:

1. Select the layer with the molecular marker data and go to *Analysis/Point to Grid/Richness*. Under *Output Variable*, select *Rarefaction*. Use the same raster file as in the richness analysis for one (1) degree cells (see Individual Task).
2. Under the *Parameters* tab, mark the SSRs using a *Standardized Sample Size* of 80.



As can be observed, the difference in diversity between northern Peru and Bolivia (where much sampling was carried out) becomes even more evident in this analysis.

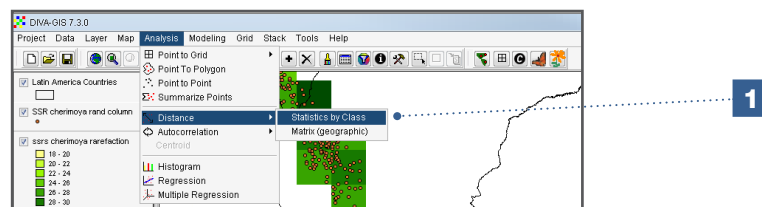
Geographical distribution of individual alleles

The previous analysis focused on the total number of alleles. Now, we will look at the distribution of individual alleles. Based on their frequencies and geographic distributions, different types of alleles can be identified. Among them, locally common alleles are the most important from a conservation standpoint as they can indicate adaptation to local conditions and are due to their restricted distribution more vulnerable to losses than broadly distributed alleles. For more information on different types of alleles, refer to Frankel et al. (1995).

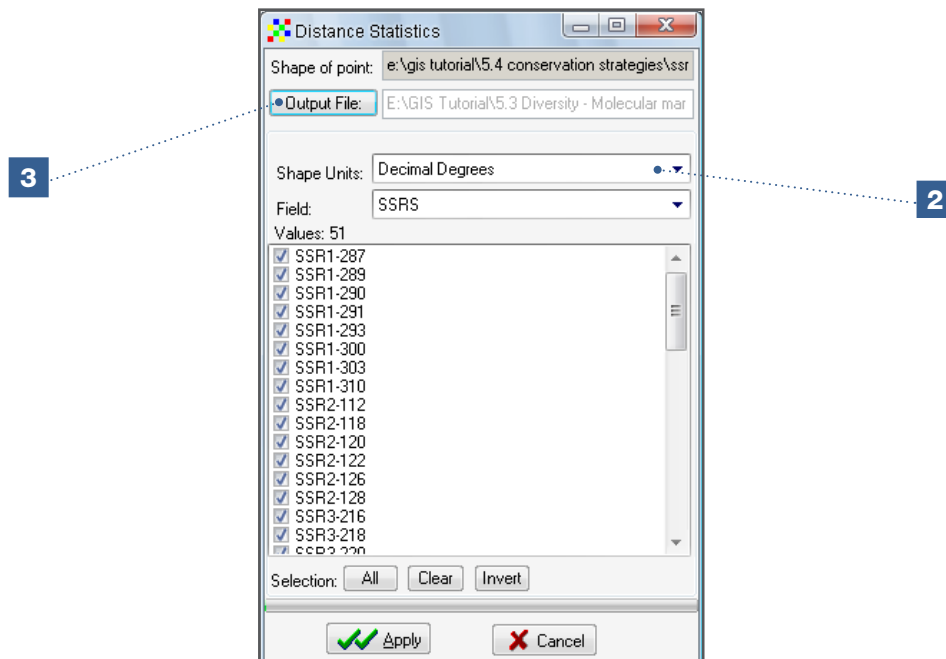
Frequency	Distribution
Frequent (5-10 %)	Local
Rare	Local
Frequent	Broad
Rare	Broad

Steps:

1. Select the layer with the molecular marker data. Go to *Analysis/Distance/Statistics by Class*.



2. Select the SSRs to calculate geographical distances.
3. Assign a name to the dBase IV file (*.dbf) that will be generated. Results can also be saved in a text file (*.txt).



CLASS	N	MAXD	MIND	AVGD
SSR1-287	2	112315.528031	112315.528031	112315.528031
SSR1-289	10	98936.601656	0.000000	51763.223028
SSR1-290	1	0.000000	0.000000	0.000000
SSR1-291	89	1182386.541258	0.000000	214634.096226
SSR1-293	923	2657272.248067	0.000000	900183.573709
SSR1-300	23	567875.621069	0.000000	202393.615713
SSR1-303	1	0.000000	0.000000	0.000000
SSR1-310	379	2643314.747000	0.000000	893601.718703
SSR2-112	169	2232607.340967	0.000000	392918.292471
SSR2-118	23	1003586.749579	0.000000	324706.614873
SSR2-120	224	2379663.891653	0.000000	616187.139674
SSR2-122	607	2657272.248067	0.000000	929194.775737
SSR2-126	8	536421.608697	0.000000	279173.418610
SSR2-129	397	2657272.248067	0.000000	926758.182019
SSR3-216	236	2634522.894471	0.000000	702397.685715
SSR3-218	567	2601975.255826	0.000000	834741.331538
SSR3-220	31	1191918.073621	0.000000	295162.040476
SSR3-222	297	2656410.934834	0.000000	852268.399690
SSR3-230	306	2601179.904615	0.000000	850633.865704
SSR3-300	1	0.000000	0.000000	0.000000
SSR4-143	48	1089564.155549	0.000000	243369.403381
SSR4-145	67	1233690.030137	0.000000	291750.047439
SSR4-147	1	0.000000	0.000000	0.000000
SSR4-152	306	2642462.654205	0.000000	943903.352155
SSR4-154	952	2657272.248067	0.000000	894483.741641
SSR4-158	55	2226441.909032	0.000000	334092.460561
SSR5-146	101	1294374.193262	0.000000	371729.455237
SSR5-156	551	2596570.336769	0.000000	890802.231982
SSR5-160	225	2530169.786344	0.000000	895319.112602
SSR5-163	63	846556.684103	0.000000	208000.229928
SSR5-164	114	1337294.585461	0.000000	363813.233640
SSR5-166	81	1176946.540977	0.000000	256772.118712
SSR5-182	35	962460.889177	0.000000	272437.476582
SSR5-184	158	2643594.122515	0.000000	676115.378699
SSR6-134	284	2657272.248067	0.000000	581088.314935
SSR6-142	410	2657272.248067	0.000000	953164.760548
SSR6-146	186	2349079.386966	0.000000	690893.442616
SSR6-149	546	2641113.593865	0.000000	864110.148842
SSR6-150	2	89782.850317	89782.850317	89782.850317
SSR7-177	53	1221947.037797	0.000000	386282.177102
SSR7-183	936	2657272.248067	0.000000	931986.969695
SSR7-189	280	2601975.255826	0.000000	959186.361269
SSR7-191	7	891200.949319	0.000000	287799.512669
SSR7-193	89	2157877.288490	0.000000	617039.274378
SSR7-206	43	1250710.467133	0.000000	311401.455394
SSR7-208	20	268120.411937	0.000000	74017.215459
SSR8-299	416	2657272.248067	0.000000	646238.190875
SSR8-301	187	2362000.788622	0.000000	853069.513845
SSR8-305	383	2568008.602972	0.000000	827545.456845
SSR8-316	467	2629913.073679	0.000000	831603.663624

When you open the dBase IV file (*.dbf) or text file (*.txt) in Excel, a summary of the number of times each allele occurs (N) and the maximum distance between alleles (MAXD) will be shown. This distance provides you with an idea of the geographic area covered by each allele. The remaining parameters, minimum distance (MIND) and average distance (AVGD), are less important in this instance.

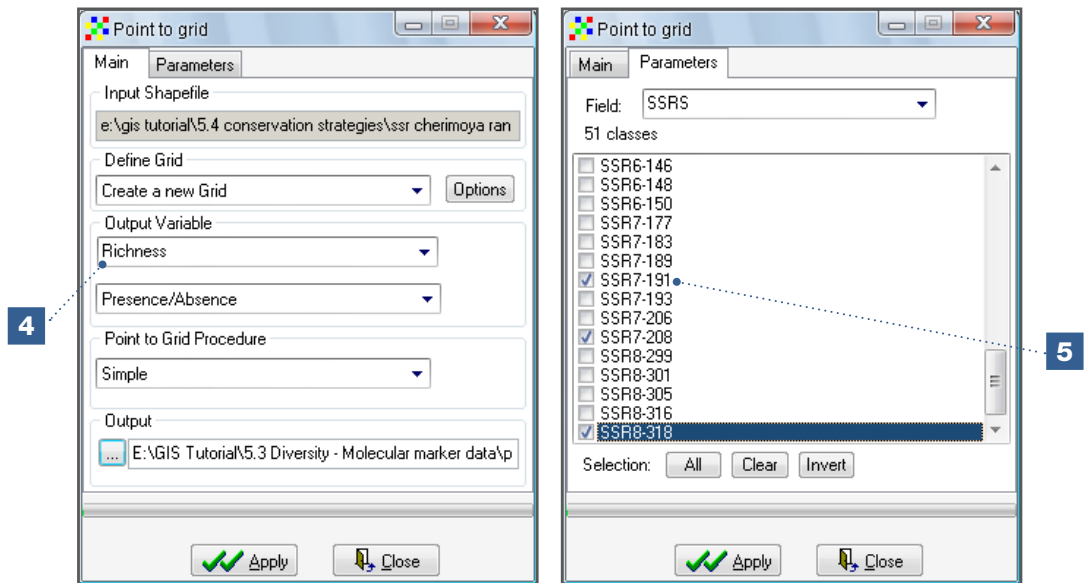
To visualize the information presented in Excel in a more organized format, change the distances from meters to kilometres, eliminate decimals and arrange the data in ascending order of MaxD. Different allele classes will then be revealed:

SSRs	N	MaxD(km)
SSR1-290	1	0
SSR1-303	1	0
SSR3-300	1	0
SSR4-147	1	0
SSR6-150	2	90
SSR1-289	10	99
SSR1-287	2	112
SSR7-208	20	268
SSR2-126	8	536
SSR1-300	23	568
SSR5-163	63	847
SSR7-191	7	891
SSR5-182	35	962
SSR2-118	23	1004
SSR4-143	48	1090
SSR5-166	81	1177

SSR1-291	89	1182
SSR3-220	31	1192
SSR7-177	53	1222
SSR4-145	67	1234
SSR7-206	43	1251
SSR5-146	101	1294
SSR5-164	114	1337
SSR8-318	5	1646
SSR7-193	89	2158
SSR4-156	55	2226
<i>Incomplete table</i>		

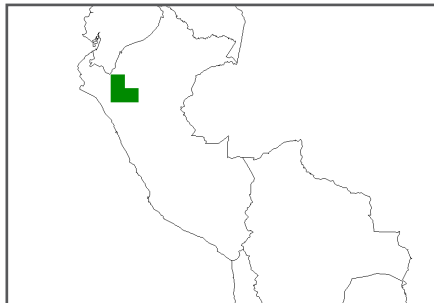
- The first group includes those alleles that have been observed only once (highlighted in yellow). These alleles are defined as unique although their occurrence depends greatly on the number of samples taken in an area. Because of the strong influence of sampling intensity on identifying unique alleles, this type of allele is considered to be of less interest than those in the following group.
 - The alleles of most interest for identifying zones of high or unique diversity are those repeatedly observed in a small area: for example, alleles SSR1-289 and SSR7-208 (highlighted in green). These are referred to as locally common alleles. The probability that this phenomenon occurs repeatedly and within a limited area is much lower than the probability of observing single alleles only once, given the differences in the sampling strategy. Populations that contain a significant number of locally common alleles may be prioritized for conservation because they contain alleles (and most likely useful genes as well) not found elsewhere, while their observation is less dependent on differences in sampling intensity.
 - Another group of alleles includes those found in a relatively large area but with very low frequencies (e.g. SSR7-191 and SSR8-318, highlighted in orange). These may indicate incomplete sampling.
 - Finally, the majority of alleles are common throughout the entire study area. These are called common alleles and do not contribute much information to a diversity analysis.
4. Look closer at the distribution of the alleles of interest: SSR1-289, SSR7-191, SSR7-208 and SSR8-318. The first step is to select the layer of the SSRs under *Analysis/Point to Grid/Richness*. This time, select *Presence/Absence* as the *Output Variable*. Use rasters with one (1) degree cells.

5. On the *Point to Grid* tab, select only the SSRs of interest (see Step 4).



Contrary to most analyses conducted in DIVA-GIS, results of the *Presence/Absence Analysis* do not display automatically. To visualize the results, open the layer and go to the sub-directory marked with the name assigned to the file. Results will be displayed in the following way:

Case 1:
Common alleles in a reduced area



SSR1-289
(10 observations, Max. distance 99 km)

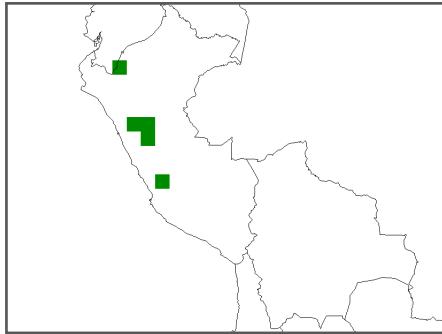


SSR7-208
(20 observations, Max. distance 268 km)

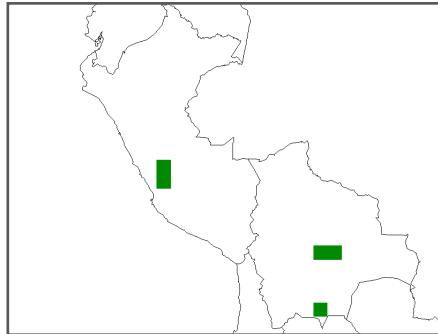
Comparing these two maps with the map of diversity generated in the previous analysis reveals that locally common alleles are found in the zones with higher richness of alleles, confirming what was previously observed: northern Peru is a hotspot for the diversity of cherimoya.

Case 2:

Rare alleles (SSR7-191 and SSR8-318) with a broad distribution



SSR7-191
(7 observations, Max. distance 891 km)



SSR8-318
(5 observations, Max. distance 1646 km)

Allele SSR7-191 is found in different areas in northern Peru but is not located within the area of greatest diversity (where the presence of unique alleles indicates rather complete sampling), suggesting that this situation is not the result of incomplete sampling. However, it may be attributed to the rules of probability, whereby (although the likelihood is low) within a random sampling scheme with sufficient observations, this situation will occur (comparable to tossing a coin several times but never having it land heads-up). Another explanation for the atypical occurrence of this allele may be the presence of local adaptation, as allele SSR7-191 is located at the same locus as allele SSR7-208 (see map above), locally common in the centre of diversity. The dominant presence of allele SSR7-208, replacing allele SSR7-191, may indicate a process of local adaptation. However, as the analysis is based on neutral markers (most likely not accounting for gene expression) such an explanation should be made with caution; in order to understand the peculiar distribution of alleles at this locus further research is required.

The situation with allele SSR8-318 is unique. The area where this allele is found covers central Peru and Bolivia; thus, there is a high probability that the allele is also present in the area between these two locations which was not sampled, i.e. southern Peru. Presence in this area would mean the allele is more common than what is indicated on the map. As such, the result suggests that the sampling was incomplete. It also indicates that there may be new alleles outside the area of highest diversity (northern Peru). As discussed in Section 5.4, these considerations are important and useful when formulating conservation strategies.

Cluster analysis

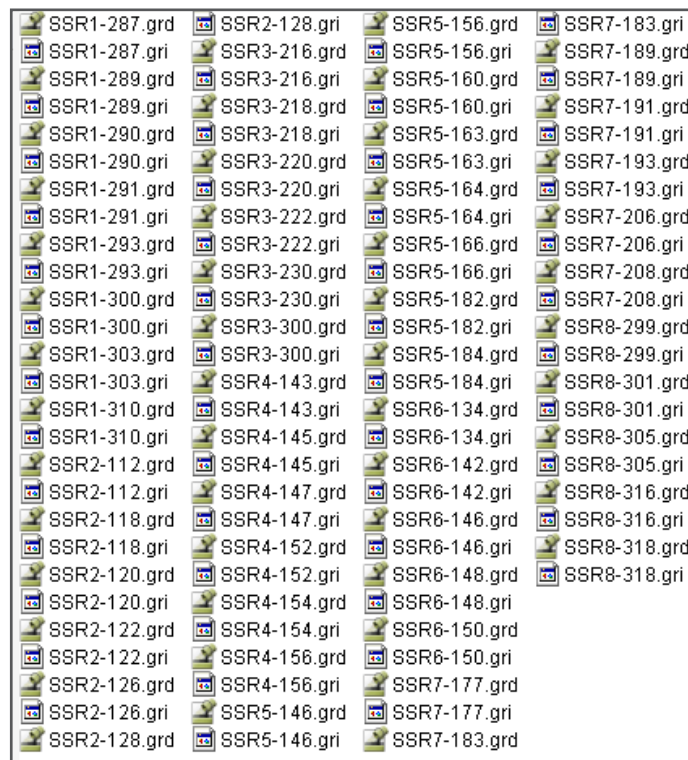
Richness analyses carried out thus far take into account the number of alleles in a raster cell (*alpha* diversity), but they disregard the variations in composition among the different cells (*beta* diversity). While two cells may have a similar richness, they may also display a completely different composition of alleles. How can diversity be analysed in this case? The DIVA-GIS programme includes a cluster analysis tool to assess differences in diversity between raster cells. In the analysis outlined below, this tool is applied in order to analyze differences in allelic composition, thereby providing further insight in the genetic structure existing across the species' geographical distribution.

Several software programmes have been developed to carry out the analysis of allelic composition in different populations. One such programme is *Structure* (<http://pritch.bsd.uchicago.edu/structure.html>), which assigns genotypes to groups based solely on allelic frequencies, independent of the subunit of the study area (raster cells) in which they are

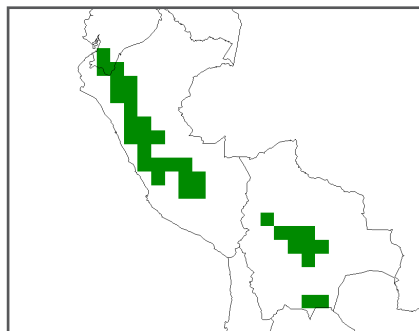
located or any other *a priori* definition of populations. For more information about *Structure* please refer to Pritchard et al. (2000). The following analysis, however, only explains how to undertake a cluster analysis in DIVA-GIS using alleles as the observed unit of diversity.

Steps:

1. Mapping different alleles is the basis for a spatial analysis of allelic composition. Repeat the mapping process done in Steps 4 and 5 of the previous section (Geographical distribution of individual alleles - *Analysis/Point to Grid/Richness with Presence/Absence*); do not select any specific alleles. Make a raster showing the distribution of each allele, using one (1) degree cells. The result generated can again be found in the newly created folder, which contains distribution rasters for each of the 51 alleles.



A few examples:

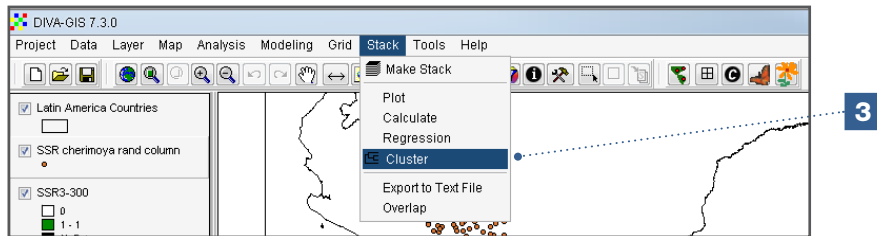


A common allele: SSR1-310

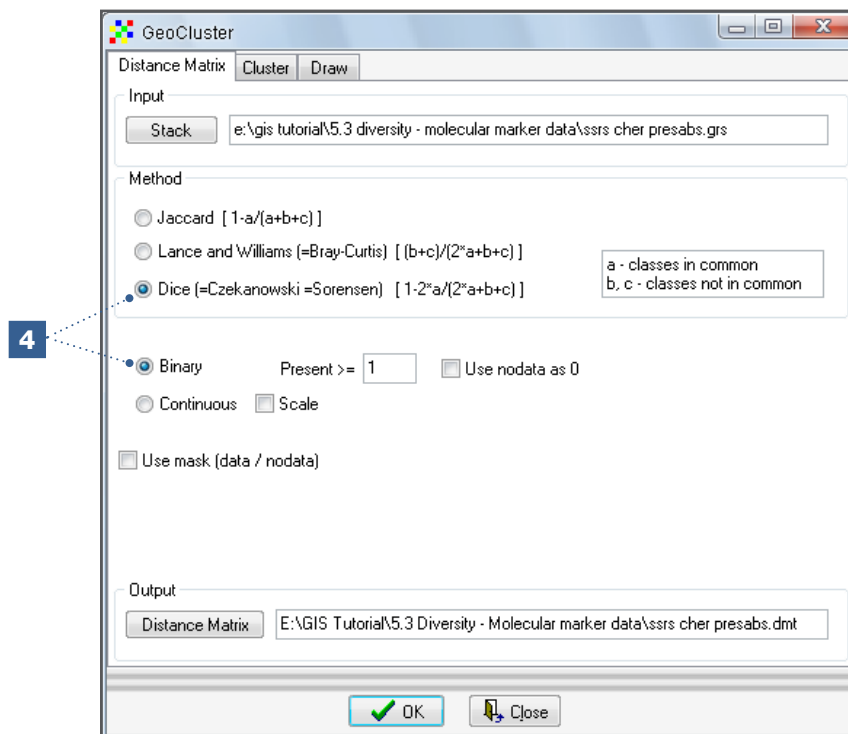


A rare allele: SSR3-300

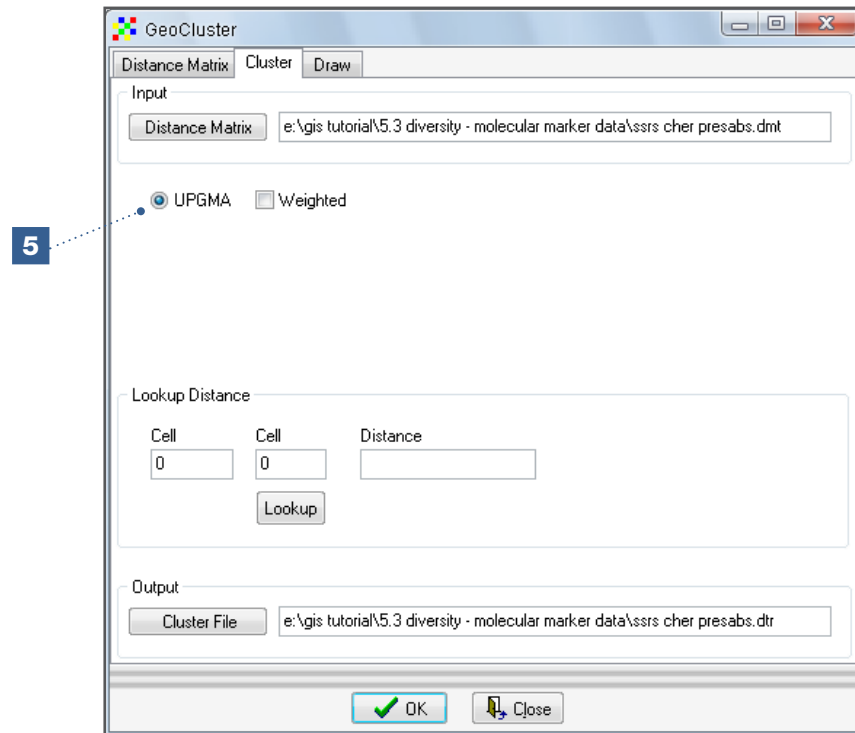
2. Before starting the cluster analysis, make a stack of presence/absence rasters for the 51 alleles generated in the previous step (see Analysis 5.2.1 Step 8). This file corresponds to a stack (group of rasters with the same properties) and will be the basis for the following analysis.
3. Go to *Stack/Cluster* and select the stack that you just created.



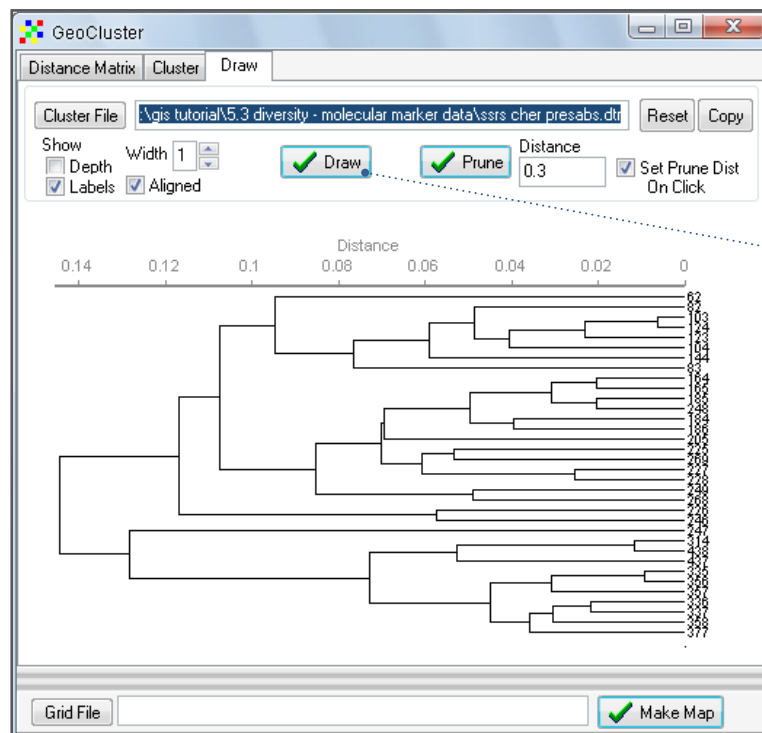
4. The next step consists of calculating the distance between different raster cells based on the composition of alleles found in each cell. Since the data being used is binary [values of one (1) for presence of one allele and zero (0) for absence of alleles], leave the default option (*Binary*) selected. Three methods are used to calculate distances. In this analysis, the *Dice* method will be utilized to calculate the matrix of distances. It is also known as the *Nei-Li coefficient* and can be used with co-dominant marker data such as microsatellite marker data (see De Vicente et al. 2004b). It is recommended to use the same filename as was assigned to the stack file. Click OK to continue.



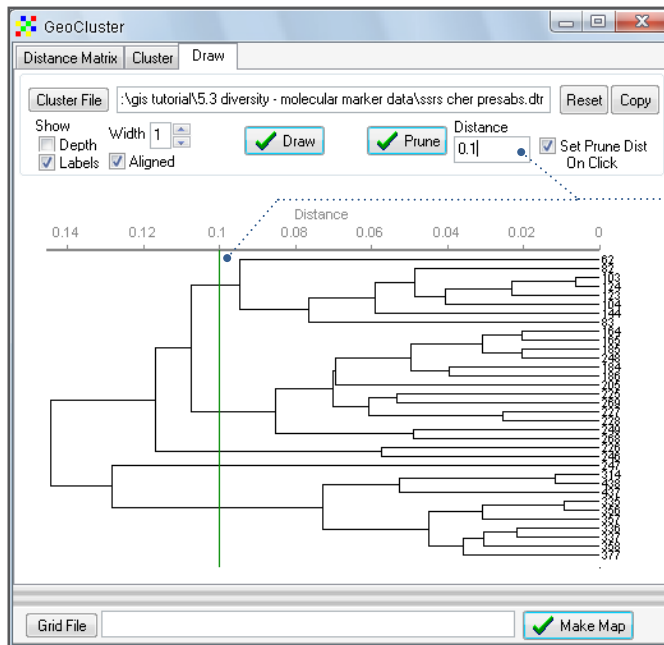
5. The next step consists of undertaking a cluster analysis based on grouping cells separated by the smallest distances (i.e. with the most similar allelic composition). Use the *UPGMA* method (default) to group the cells. Click *OK* to continue (again, it is recommended to use the same file name).



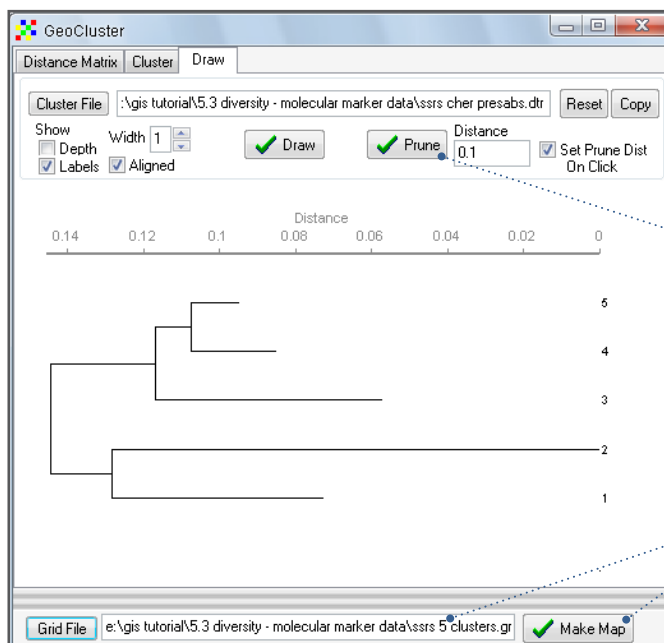
6. Under the *Draw* tab, click on the *Draw* button. A dendrogram (tree) will be displayed showing similar cells, i.e. those with a similar allelic composition.

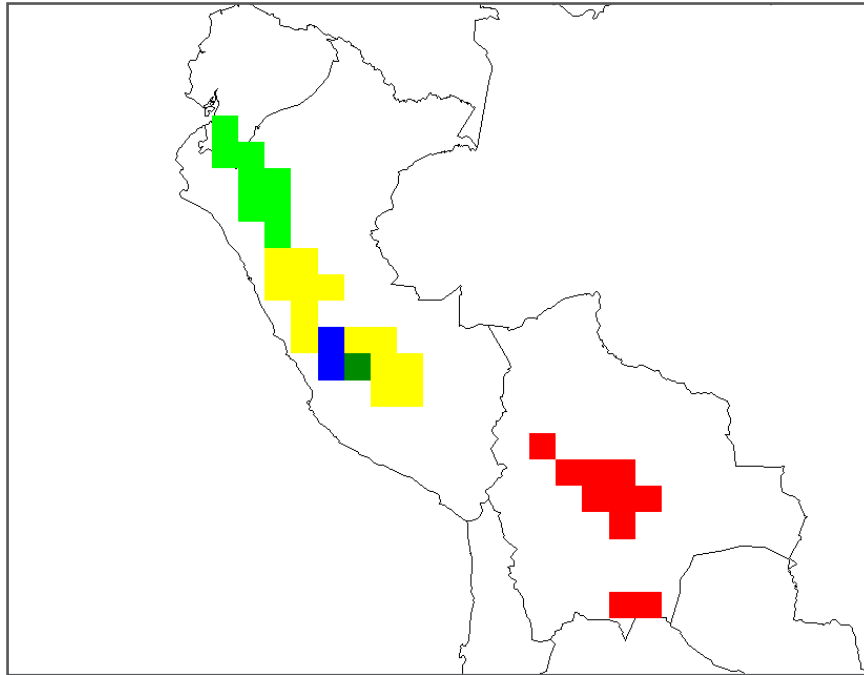


- Cells can be grouped according to the level of similarity (as indicated by small distances). This can be done by pruning the dendrogram. Click on the dendrogram at the distance value where you wish to group the cells and then click the *Prune* button (for example, at a distance of 0.1). Alternatively, you can enter the desired value for clustering the cells in the *Distance* window.



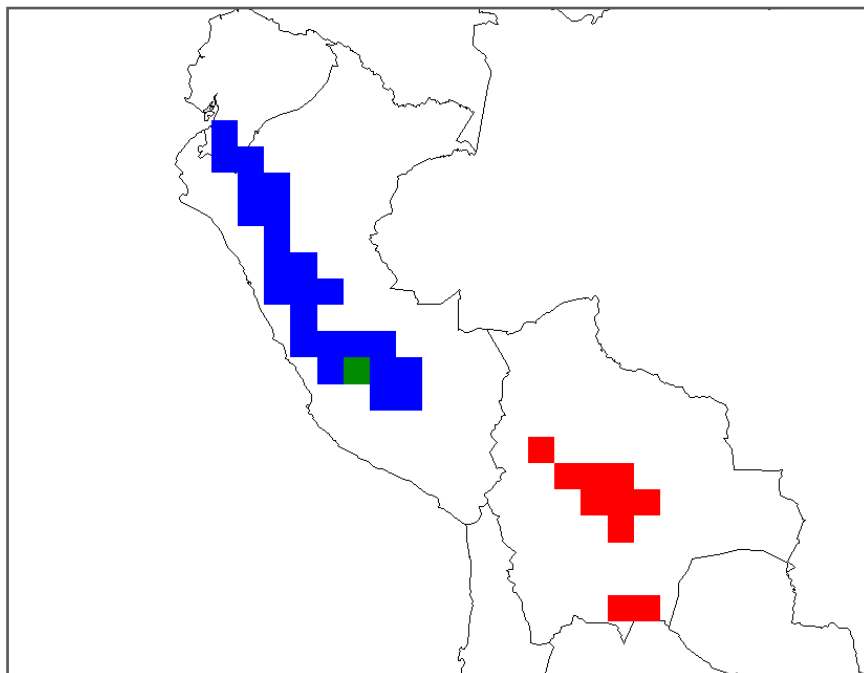
- Click on the *Prune* button to cut the tree at the selected distance. The resulting grouping of similar cells will be displayed as a tree with fewer branches (the complete dendrogram can be displayed again using the *Draw* button). When pruning at a value of 0.1, five groups will remain.
- To visualize the result, assign a name to the raster in the *Grid File* window and then click on *Make Map*.





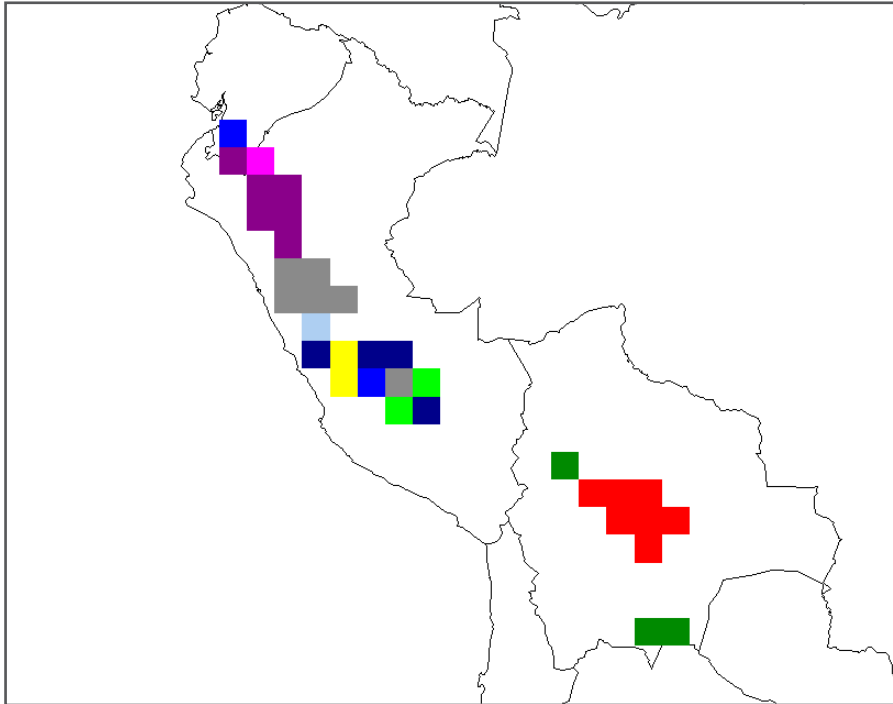
As a result of grouping similar cells in five different groups, three large zones can be observed, each having a different allelic composition: one zone covers northern Peru and southern Ecuador; the second covers the central part of Peru; and the third is located in Bolivia.

The genetic structure of cherimoya distribution can be further explored using different pruning points. However, be aware that differences may no longer be apparent when the number of groups is reduced. For example, when cutting at a distance of 0.12, only three groups are left.



The separation between most populations in Peru and Ecuador has disappeared as a result of the cutting. On the other hand, the cherimoya accessions in Bolivia are confirmed to be different from those in Ecuador and Peru. Although the previous analyses indicated Bolivia had low levels of diversity, the allelic composition of the materials in the country is quite distinct from that of the other study areas.

- When there are too many groups, it is also difficult to identify patterns in the composition of alleles. Cutting the dendrogram at a distance of 0.065 generates 11 groups of alleles.



Having too many groups, as illustrated in the map above, complicates the interpretation of results. Despite the high number of groups, the accessions in Bolivia continue to represent a homogenous group, separated from the others.

5.4. Implications for the formulation of conservation strategies

The three previous analyses focused on using spatial analyses to detect areas of high diversity (*alpha* diversity) and, to a lesser extent, to understand differences in the diversity between areas (*beta* diversity). An understanding of the extent and distribution of diversity is critical to designing effective and appropriate conservation strategies. It is also vital in order to identify key sites for carrying out *ex situ* and *in situ* conservation activities (priority areas for collection and protection), particularly since resources allocated for conservation are frequently scarce, limiting the possibility to conduct conservation actions in several areas. Combining *alpha* and *beta* diversity analyses allows for the optimization of resources to conserve the greatest amount of diversity possible. An area with the highest *alpha* diversity is usually prioritized as the first site for *in situ* conservation (excluding any logistical constraints). However, a more complex question exists as to

what area should be considered as the second priority, if additional resources remain available. The second priority area should not necessarily be that with the second highest degree of *alpha* diversity, as a large portion of this diversity may already be conserved in the first priority area. In this case, *beta* diversity must be taken into account, focusing on areas with a species or allelic composition different from that of the area already prioritized.

This concept of complementarity is considered in DIVA-GIS. Conservation activities generally focus on including the highest number of species but may also target the conservation of a particular species. In this case, alleles are used as the observed unit of diversity to define priority *in situ* gene conservation areas. This concept is outlined in the following analysis focusing on cherimoya.

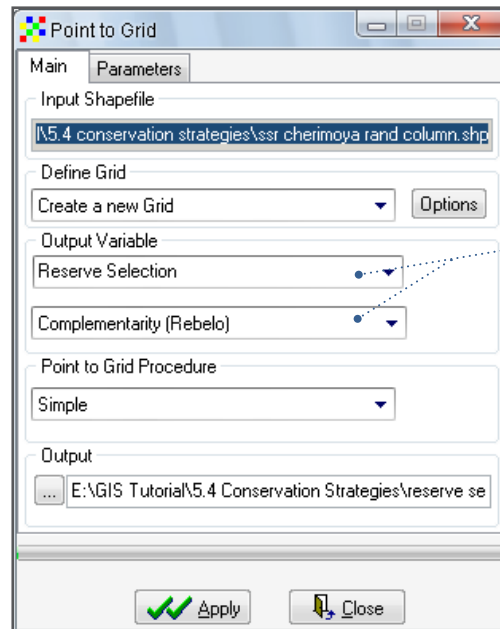
PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel 	<p>Data Files:</p> <p>Folder 5.4 Conservation Strategies</p> <ul style="list-style-type: none"> • <i>SSR cherimoya rand column (shp, shx, dbf)</i> • <i>Latin America Countries (shp, shx, dbf)</i> • <i>Protected Areas Latin America (shp, shx, dbf)</i>

5.4.1. How to identify priority zones for *in situ* conservation or germplasm collection

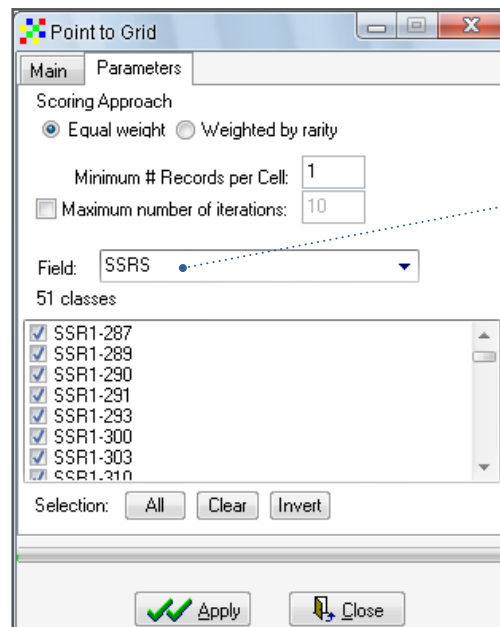
Through this analysis, you will learn to use the *Analysis (Point to Grid and Reserve Selection)* option in DIVA-GIS to assist in defining priority conservation areas. The *Reserve Selection* procedure uses an optimization algorithm that was originally developed to minimize the area needed to conserve flowering plant diversity in South Africa (See Rebelo and Siegfried 1992). In this analysis, we will use that algorithm to define the minimum number of geographic units needed to conserve all genetic diversity (measured through molecular markers) and to identify, in sequence of importance, the geographic units that should be prioritized for conservation. The data used in the previous analysis will be used again. The same type of analysis can be run at the species level, as well.

Steps:

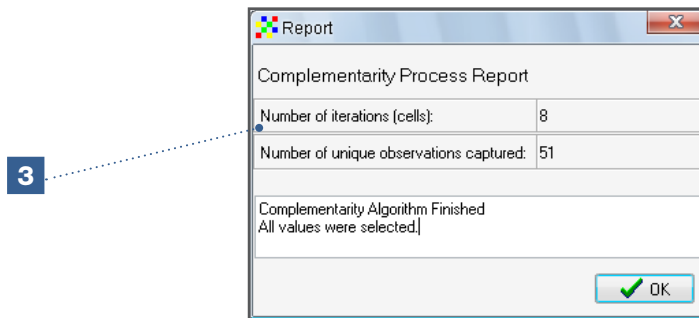
1. Use the molecular marker data for cherimoya to start the analysis and select the *Analysis/Point to Grid* option. For this analysis, select the *Reserve Selection* option, as well as the *Complementarity* option. Use a raster with one (1) degree cells (preferably one of the rasters used in the previous analyses).



2. Each allele should receive the same weight (default option), yet you can also give more weight to rare alleles or species, if necessary.

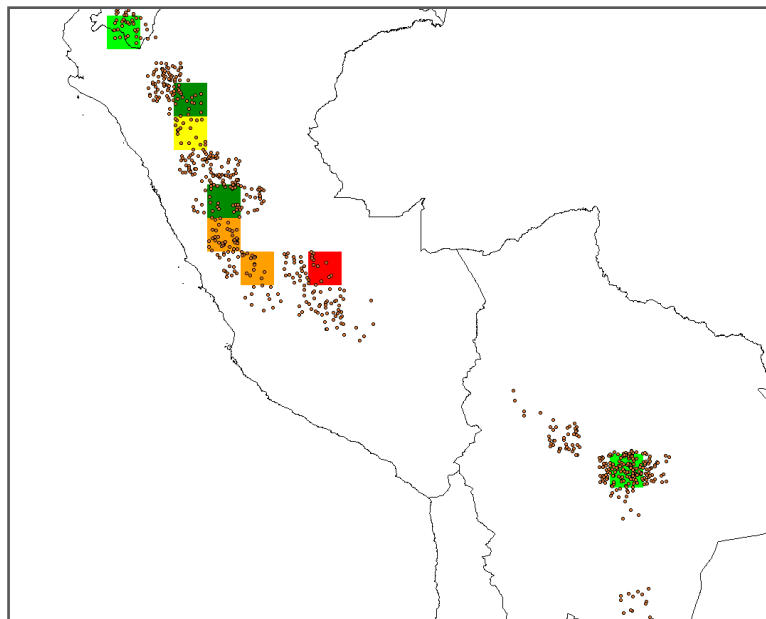


3. The first output visualized is a report on the number of cells selected (eight) necessary to conserve all alleles and the number of unique observations (the 51 alleles) generated by the analysis.



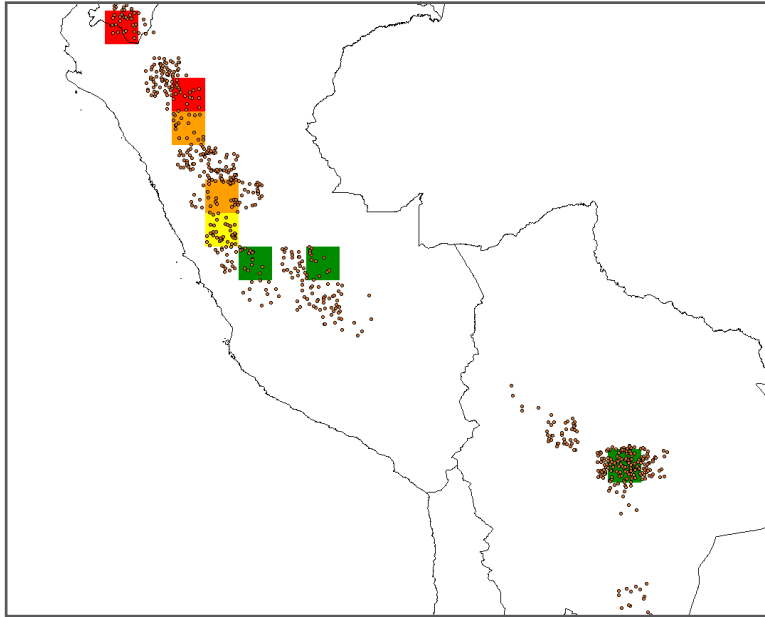
4. Next, click *OK* on the *Report* window. Three rasters titled *Sequence*, *Classes* and *Additional Classes* will be displayed. Before visualizing each result, improve the legend using the *NoData transparent* option to broaden the number of classes.

- *Sequence*



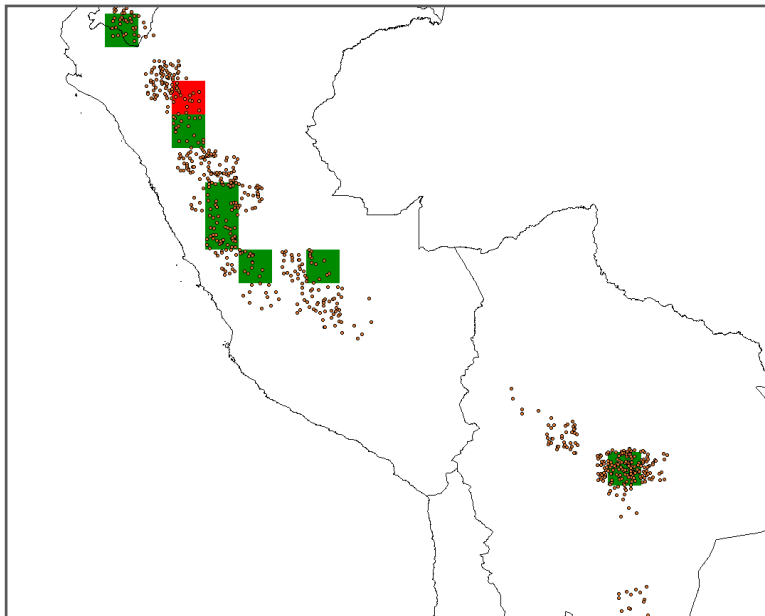
This raster corresponds to the most important result: the sequence of cell selection. In other words, the raster indicates the priority cells for conservation [the cell with a value of one (1) is the most important]. By selecting the layer and locating the pointer over each cell, the values can be seen in the lower part of the map, on the status bar. Here, the priority cells in order of importance include: 1) the cell located in northern Peru; 2) the cell located a bit further south; 3) the cell in Bolivia; and 4) the cell located in the southern part of Ecuador. This analysis highlights that cells were not selected based only on the highest diversity, but also based on differences in allelic composition (as illustrated by including a cell from Bolivia, for which diversity is not very high).

- *Classes*



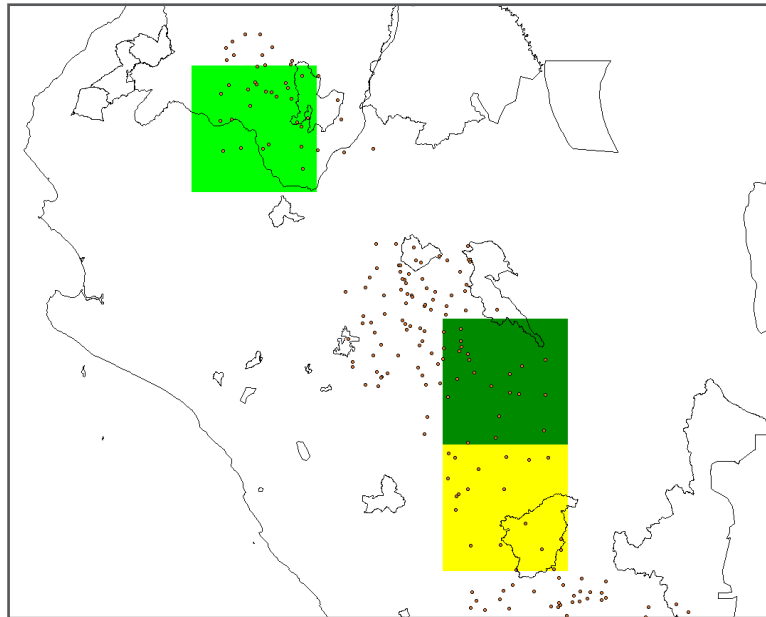
This raster corresponds to a richness analysis, showing only the information (richness) for those cells selected in the previous analysis. Using the pointer, it is possible to verify that the cell selected as having the highest priority actually corresponds to a cell with the highest level of diversity in the group (39 alleles). Note: in this analysis, the second priority cell is not the one with the second highest level of diversity.

- *Additional Classes*



This raster reveals the number of new alleles contributed by each additional cell, taking into account that all cells selected cover the 51 alleles. The first cell contributes the largest number of alleles (39 alleles).

5. To complete the analysis, add the layer of protected areas (using the *Protected Areas Latin America.shp* file) and analyze the current status of conservation for priority cells based on the Sequence layer.



This final step reveals that there are protected areas in each one of the cells, providing the first indication of *in situ* conservation of cherimoya (more detailed analyses are required for real situations). The results of the analysis seem to indicate that cherimoya diversity is currently partially conserved since protected areas are located in all priority cells. The cell with the highest priority, however, has low coverage in a protected area. An analysis at a smaller scale would enable more in-depth conclusions by showing whether cherimoya accessions with important alleles are actually included in these protected areas.

References

- Chapman AD 2005. Principles of Data Quality, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. [on line] Available from: <http://www2.gbif.org/DataQuality.pdf>. Date accessed: October 2010.
- De Vicente MC, Lopez C, Fulton T. 2004a. Molecular marker learning module volume 1: Using Molecular Marker Technology in Studies on Plant Genetic Diversity: Learning module. IPGRI and Cornell University.
- De Vicente MC, Lopez C, Fulton T. 2004b. Molecular marker learning module volume 2: Genetic diversity analysis with molecular marker data: Learning module. IPGRI and Cornell University.
- Frankel OH, Brown AHD, Burdon JJ. 1995. The conservation of plant biodiversity. Cambridge University Press, Cambridge, UK.

- Grum M, Atieno F. 2007. Statistical analysis for plant genetic resources: clustering and indices in R made simple. Handbooks for Genebanks, No. 9. Bioversity International, Rome, Italy.
- Hajeer A, Worthington J, John S, editors. 2000. SNP and microsatellite genotyping: Markers for genetic analysis. In: Biotechniques: Molecular laboratory methods series. Eaton Publishing, Manchester, UK.
- Hijmans RJ, Garrett KA, Huamán Z, Zhang DP, Schreuder M, Bonierbale M. 2000. Assessing the geographic representativeness of genebank collections: the case of Bolivian wild potatoes. *Conservation Biology* 14(6): 1755-1765.
- Kindt R, Coe R. 2005. Tree diversity analysis. A manual and software for common statistical methods for ecological and biodiversity studies. World Agroforestry Centre (ICRAF), Nairobi.
- Leberg PL. 2002. Estimating allelic richness: effects of sample size and bottlenecks. *Molecular Ecology* 11: 2445-2449.
- Mathur PN, Muralidharan K, Parthasarathy VA, Batugal P, Bonnot F. 2008. Data Analysis Manual for Coconut Researchers. Bioversity Technical Bulletin No. 14. Bioversity International, Rome, Italy.
- Petit RJ, El Mousadik A, Pons O. 1998. Identifying populations for conservation on the basis of genetic markers. *Conservation Biology* 12: 844-855.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945-959. [on line] Available from: <http://pritch.bsd.uchicago.edu/publications/structure.pdf>. Date accessed: October 2010.
- Rebelo AG, Sigfried WG. 1992. Where should nature reserves be located in the Cape Floristic Region, South Africa? Models for the spatial configuration of a reserve network aimed at maximizing the protection of diversity. *Conservation Biology* 6 (2): 243-252.
- Scheldeman X, Willemen L, Coppens D'eeckenbrugge G, Romeijn-Peeters E, Restrepo MT, Romero Motoche J, Jimenez D, Lobo M, Medina CI, Reyes C, Rodriguez D, Ocampo JA, Van Damme P, Goetghebeur P. 2007. Distribution, diversity and environmental adaptation of highland papaya (*Vasconcellea* spp.) in tropical and subtropical America. *Biodiversity and Conservation* 16(6): 1867-1884.
- Willemen L, Scheldeman X, Soto Cabellos V, Salazar SR, Guarino L. 2007. Spatial patterns of diversity and genetic erosion of traditional cassava (*Manihot esculenta* Crantz) cultivation in the Peruvian Amazon: an evaluation of socio-economic and environmental indicators. *Genetic Resources and Crop Evolution* 54(7): 1599-1612.

Chapter 6

Species distribution modelling and analysis

Ecological niche is a theoretical concept frequently used in biodiversity conservation studies. The concept has been successfully applied to prioritize sites for the *in situ* conservation of wild species and in planning collection missions for crop genetic resources (e.g. Jarvis et al. 2005; Scheldeman et al. 2007).

While different definitions for ecological niche have been formulated, the concept basically refers to the environmental space a species occupies under natural conditions (Puliam 2000). The generally accepted definition, as provided by Hutchinson (1957), distinguishes between a fundamental niche and a realized niche. A fundamental niche is the range of environmental conditions under which a species can theoretically exist, whereas the realized niche is defined by the combination of negative interactions (e.g. competition and predation) that restrict a species' presence and positive interactions (e.g. facilitation) that expand the environmental ranges in which a species is able to grow.

Geographic information systems (GIS) such as DIVA-GIS include the ability to model ecological niches based on available environmental data from sites where species have been observed (presence points). Databases which provide detailed climatic data (based on interpolations of data collected by climatic stations worldwide) already exist, such as Worldclim (Hijmans et al. 2005), but the availability of data for other relevant environmental factors, such as soil variables, is still limited. Thus, many GIS tools approximate the value of the ecological niche using climatic variables known as the 'climate envelope' (Guarino et al. 2002).

Analyses of the ecological niche of wild species and their genetic resources using presence points are applied under several conditions, the most important being:

1. The species should be in a state of equilibrium with its environment; in other words, the environmental ranges are restricted by competition and predation, and not by dispersion limitations.
2. The available environmental variables (e.g. climate variables) used in the modelling are determinant abiotic factors in shaping the natural species distribution.
3. No presence points should be included of specimens grown in plantations, field collections and botanical gardens which may be located in an environment outside the realized niche of the species.

In practice, one or more of these conditions above are often not met; nonetheless, species distribution modelling is still a useful tool to approximate the realized niche and the natural distribution of a species. As such, species distribution modelling is useful for prioritizing conservation activities.

The ecological niche concept can also be used to identify agroecological zones ideal for growing specific crops and trees. Ecocrop (FAO 2007), which is included in DIVA-GIS, as well as Homologue (Jones et al. 2005) are examples of modelling programmes that can be utilized for such analyses. The identification of optimum production zones is a more complex task than identifying the natural distribution area due to the dynamic interaction

between agricultural management practices and environmental factors. Under specific agricultural practices, a species can be cultivated in an environment outside its realized niche, and even its fundamental niche, if additional resources are available, including water through irrigation or soil nutrients through fertilization.

Finally, it is important not to overlook the relationship between genetic variation and the environment. Populations of wild species and crops are capable of evolving locally, adapting to site-specific environments. This results in differences in phenotypic expression between individuals of different populations, even when they are planted together at a common site. Genotypes x Environment (GxE) experiments enable one to identify the most appropriate ecotypes and varieties for a specific agroecological zone. This type of analysis is relevant for selecting promising germplasm adapted to specific areas, but will not be discussed here as it is beyond the scope of this manual.

This chapter focuses on the use of modelling to predict species distribution and how its results can contribute to the prioritization of sites for conservation, climate change impact studies and species germplasm collection.

6.1. Analysis of the realized niche of a species

As mentioned in the introduction of this chapter, several GIS software programmes, including DIVA-GIS, include simplified species distribution models based on climate data from the presence points of individuals and/or groups of individuals of a species.

If only a limited number of presence points is available for a given species, the corresponding environmental data may not adequately represent the species' realized niche. This can result in the significant misinterpretation of the results. Therefore, it is important to ensure that the presence points used in models are of good quality (see Chapter 4) and of sufficient quantity in order to obtain a sound species distribution model. It should be noted that there is no standard in terms of the minimum number of points required, as often this will relate to the nature of the species. For rare species or species with a restricted niche, only a small number of presence points may exist. However, in these cases, even a small number of points may be highly representative of the niche. As such, strict guidelines on the minimum number of presence points necessary to undertake credible species distribution modelling cannot be provided. However, a few examples of the number of presence points used for specific species are available: a) studies conducted by Scheldeman et al. (2007) used a minimum of 10 points for rare *Vasconcellea* species with a restricted distribution; b) the MAPFORGEN project - which evaluated the natural distribution of 100 species native to Latin America - used a minimum number of 20 species presence points as its threshold; and c) Van Zonneveld et al. (2009a) worked with a minimum number of 50 presence points for two pine tree species with a broad geographic distribution throughout Southeast Asia.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Excel 	<p>Data Files:</p> <p>Folder 6.1 Realized niche</p> <ul style="list-style-type: none"> • <i>Vasconcellea_4species</i> (<i>shp, shx, dbf</i>) <p>For this analysis, you need to have the 2.5 min worldclim climate data imported in DIVA (cf. Section 2.2)</p>

6.1.1. How to analyze and compare realized niches of different species

Vasconcellea species are adapted to different environments. To evaluate these differences, the climate data of the respective presence sites for each species can be extracted using DIVA-GIS (see Chapter 3). Understanding of the adaptive capacity of each species to different climatic conditions will provide information on the potential for inclusion in breeding programmes (e.g. cold tolerance in papaya) as well as for the commercial growing of *Vasconcellea* species, taking into account optimal agroecological zones. Identification of the limits of the species' niche also provides key information for conservation; for example, species that occur in narrow climate ranges are likely to be more vulnerable to climate alterations than those with a broad climatic niche.

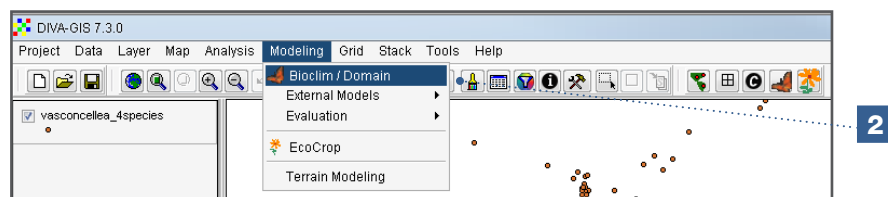
The next section outlines how to determine the niche of the following *Vasconcellea* species: *V. cundinamarcensis*, *V. microcarpa*, *V. quercifolia*, and *V. parviflora*. The section explains how to utilize the *Modeling* menu in DIVA-GIS to analyze climatic niches and to compare the niches of different species using Excel.

Using the modeling menu in DIVA-GIS to examine the realized niche of a species

Steps:

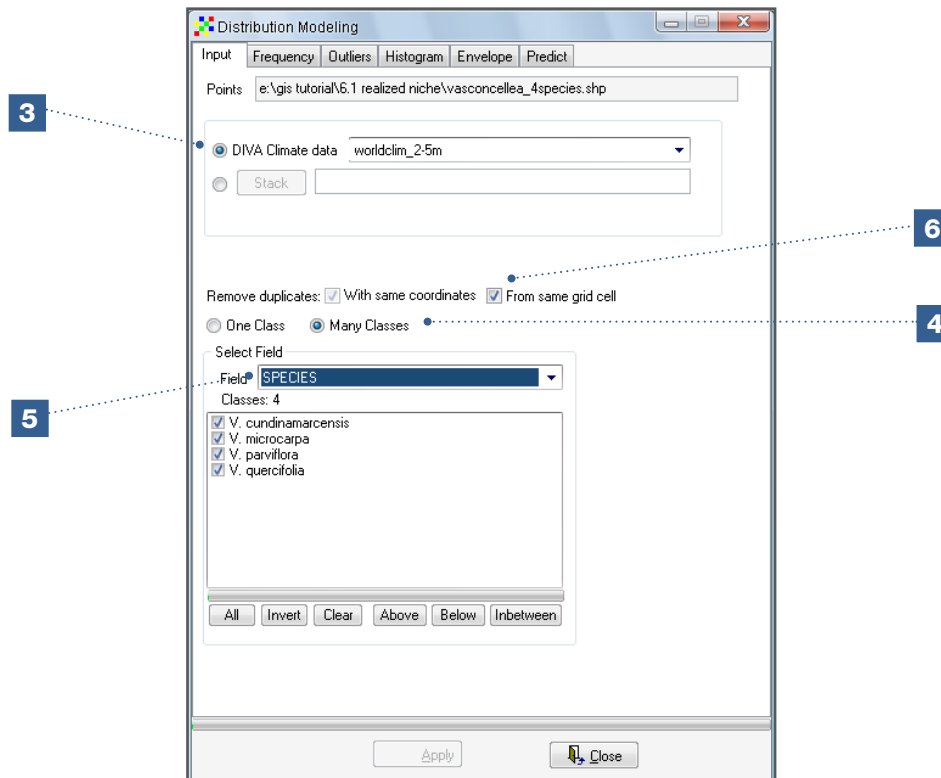
After importing the Bioclim climatic variables to DIVA-GIS (see Chapter 2), the realized niche of a species can be identified based on the species' presence points and corresponding climate data. It is possible to display the realized niche in different ways using DIVA-GIS: by visualizing the frequencies of the different climate parameter ranges as histograms; or by visualizing a two-dimensional climatic niche based on two climate parameters. These options for analysis are available in the *Modeling* menu in DIVA-GIS.

1. Open the *Vasconcellea_4species.shp* file in DIVA-GIS and keep the file as the selected layer.
2. Go to *Modeling/Bioclim/Domain* to open the *Distribution modeling* window.



3. In the *Distribution modeling* window, under the *Input* tab, go to *DIVA climate data* and select the climate database to be used. This analysis will use the Worldclim climate data at a 2.5-minute resolution (file: *wordclim_2-5min*). This was the file imported to DIVA-GIS in Chapter 2.
4. Select the *Many Classes* option to distinguish between the different classes within the vector file (*.shp), i.e. different species, genotypes, countries. This option must be selected as this analysis focuses on four different species.
5. In the *Field* window, select the parameter that will be used to define the different classes. In this case, select *Species*. The complete list of *Vasconcellea* species will be displayed.

- Indicate whether or not you wish to remove duplicate presence points within the same cell in the climate raster. For this analysis, select the *Remove duplicates: From same grid cell* option. This will prevent cells with many observations from contributing disproportionately to defining the realized niche.

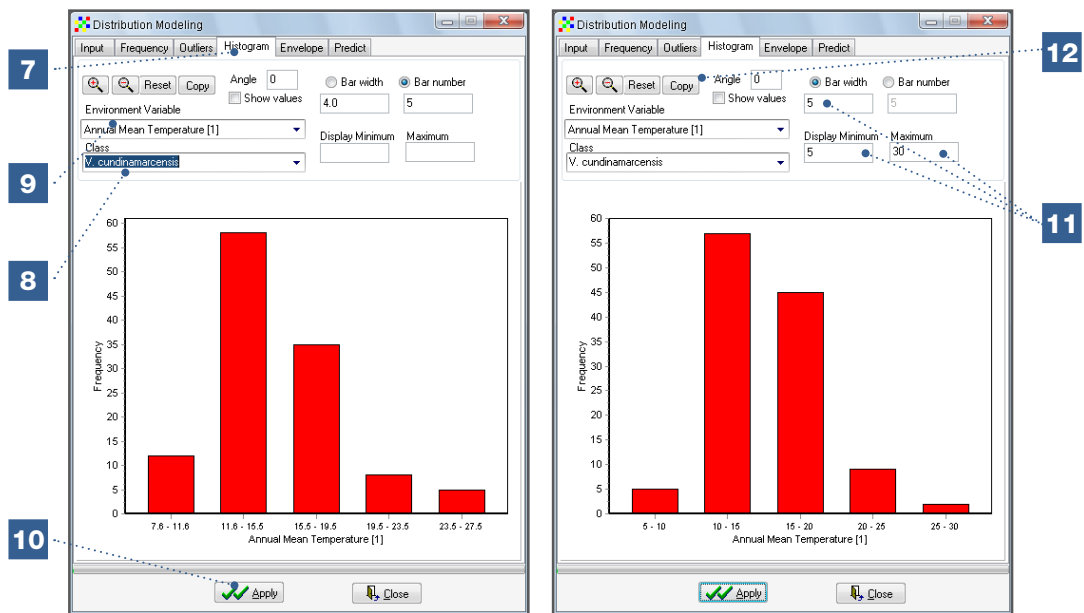


The menu includes five types of analysis (five different tabs), two of which will be explained in this section: *Histogram* and *Envelope*. The *Frequency* and *Outliers* analysis options were already described in Chapter 4. The *Predict* option models potential species distribution using the Bioclim algorithm. In this manual, however, species distribution modelling is done with another method and software (Maxent), and the *Predict* option is therefore not explained in further detail.

Histogram

- The *Histogram* tool constructs frequency histograms, which show the distribution of a species along customized ranges for different climatic variables.
- Select the desired species to generate a histogram. For this analysis, select *V. cundinamarcensis*.
- Select the desired climatic variable to visualize the climate range frequencies where this species was observed. In this case, select *Annual Mean Temperature*.
- Click *Apply* to display the histogram.
- The width and number of bars of the histogram, as well as the maximum and minimum values, can be modified. In this analysis, 5 °C and 30 °C are defined as the minimum and maximum values in order to generate five bars at five-degree intervals.

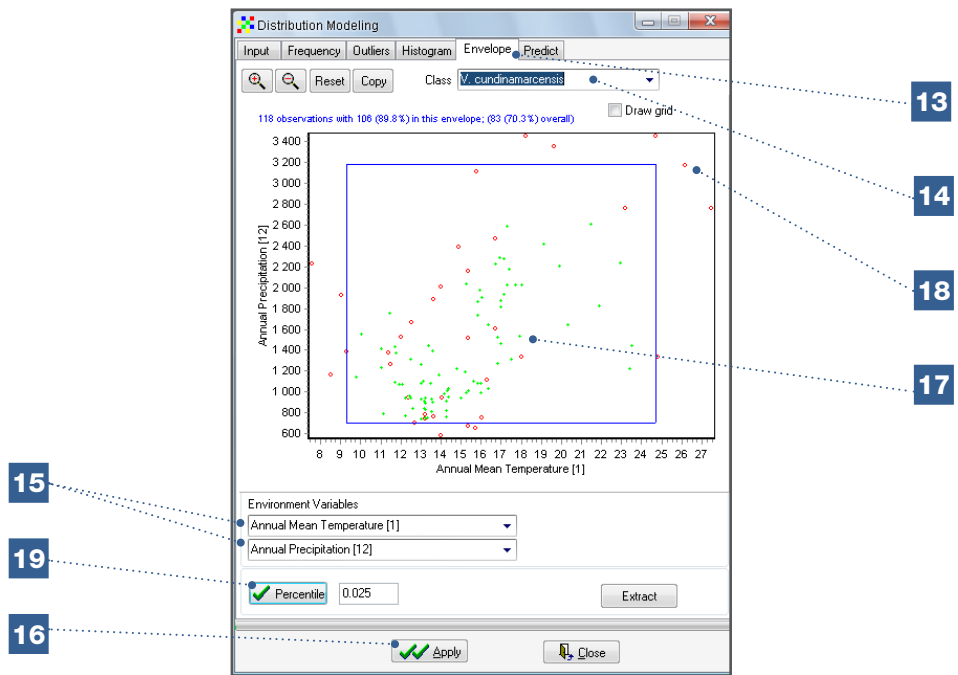
12. The histogram can be copied and pasted directly into a document using the *Copy* option.



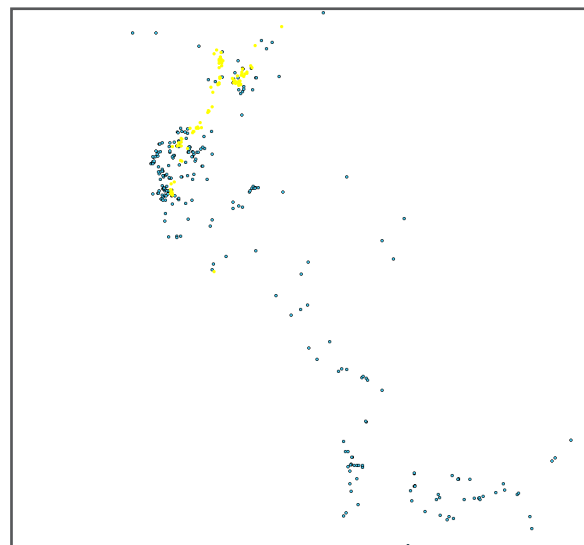
Envelope

13. The *Envelope* tool allows you to visualize a two-dimensional niche based on two climatic variables.
14. Select the desired species to generate a two-dimensional niche. Look at the different *Vasconcellea* species, particularly *V. cundinamarcensis*.
15. Select the two climatic variables on which the climatic niche will be based. For this analysis, select *Annual Mean Temperature* and *Annual Precipitation*.
16. Click *Apply* to visualize the two-dimensional niche.
17. Green points within the blue rectangle of the climatic niche represent those presence points with a climate profile within the range limits of all 19 Bioclim climatic variables.
18. Red points represent presence points with a climate profile of which one or more of the values of the 19 Bioclim climatic variables are outside the range limits. Red points within the blue rectangle are presence points having a climate profile with values for the selected variables (*Annual Mean Temperature* and *Annual Precipitation*) within the range limits of these two variables, but with values of one or more of the other 17 Bioclim variables values outside the range limits.

19. The limits of the two-dimensional niche are, by default, the 0.025 (the lowest 2.5 % of the points) and 0.975 (the highest 2.5 % of the points) percentiles, meaning that 95 % of the presence points have been taken into account in developing the two-dimensional niche (5 % are considered as outliers). The niche width can be adjusted by changing the value of the percentile. Depending on your interest, you can narrow the niche to determine the environmental ranges of the core of the species distribution or enlarge the niche to determine the extreme values under which the species can still occur (it is recommended to only do this after data quality has been checked, as atypical values can significantly influence the ecological niche when all points, percentile value 0, are included).



20. The main DIVA-GIS window will display all points (in yellow) corresponding to the green points in the *Envelope* analysis. The points remaining in the original colour (blue, in this case) correspond to points with a climate profile of which one or more of the values of the 19 Bioclim climatic variables are outside the range limits.



Comparison of the realized niches of different species in Excel

DIVA-GIS allows climate data for the corresponding presence points to be exported using the *Extract values by points* option (see Chapter 3). Data can then be analysed in Excel spreadsheets for further visualization.

Individual Task: Extract the bioclimatic climate variables from the worldclim_2-5m data (based on Data/Extract Values by Points/From Climate Data) for all presence points in the *Vasconcellea_4species.shp* file (see Analysis 3.1.4).

Steps:

- The text (*.txt) file with the extracted climate data can be opened using Excel. Each row represents a presence point. Columns BIO1, BIO2... BIO19 correspond to the 19 Bioclim variables with the values corresponding to each presence point.

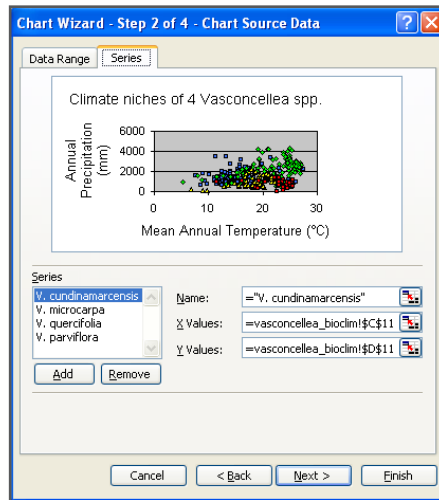
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	
ID	SPECIES	LATITUDE	LONGITUDE	COUNTRY	ADM1	RecNo	PointNo	Lon_ext	Lat_ext	bio1	bio2	bio3	bio4	bio5	bio6	bio7	bio8	bio9	bio10	bio11	bio12	bio13	bio14	bio15	bio16	bio17	bio18	bio19	
1	3433 V. cundina	6.9636	-75.4177	Colombia	Antioquia	1	1	-75.4177	6.9636	15.8	10.31667	83.87534	41.99667	21.5	9.2	12.3	15.6	15.46667											
2	3037 V. cundina	7.1716	-75.7633	Colombia	Antioquia	2	2	-75.7633	7.1716	21.49583	10.25833	83.40108	40.81323	27.7	15.4	12.3	21.23333	21.71667											
3	2816 V. cundina	7.0333	-75.3166	Colombia	Antioquia	3	3	-75.3166	7.0333	19.64583	9.775	78.83065	45.45019	25.2	12.8	12.4	19.46667	19.65	21.7										
4	2836 V. cundina	6.3333	-75.25	Colombia	Antioquia	4	4	-75.25	6.3333	17.31667	9.566667	90.25157	34.06767	22.5	11.9	10.6	17.13333	17.11667											
5	3030 V. cundina	6.9	-75.9666	Colombia	Antioquia	5	5	-75.9666	6.9	16.98333	10.2	85	41.79677	22.9	10.9	12	16.86667	17.1	12										
6	3418 V. cundina	6.1666	-75.5866	Colombia	Antioquia	6	6	-75.5866	6.1666	18.0625	9.625	89.95327	34.78015	23.6	12.9	10.7	17.75	17.86667											
7	3336 V. cundina	6.4861	-75.3952	Colombia	Antioquia	7	7	-75.3952	6.4861	17.44167	10.11667	86.46724	51.51493	23.2	11.5	11.7	17.18333	17.08333											
8	3336 V. cundina	6.2833	-75.4333	Colombia	Antioquia	8	8	-75.4333	6.2833	17.025	9.733333	89.29664	34.21191	22.4	11.4	10.9	10.76667	16.85	17										
9	3336 V. cundina	6.4627	-75.5563	Colombia	Antioquia	9	9	-75.5563	6.4627	16.39167	10.1	84.87395	50.71459	22.3	10.4	11.9	16.15	16.5	17										
10	3339 V. cundina	6.335	-75.5527	Colombia	Antioquia	10	10	-75.5527	6.335	20.36667	10.98333	88.57527	39.808	26.7	14.3	12.4	20.68333	20.1											
11	3027 V. cundina	6.8477	-75.4608	Colombia	Antioquia	11	11	-75.4608	6.8477	15.25833	10.48633	86.73611	42.983	21.1	9.1	12	15.88333	14.93333											
12	3354 V. cundina	5.8	-75.5666	Colombia	Antioquia	12	12	-75.5666	5.8	22.97917	10.675	87.5	45.79889	29.5	17.3	12.2	22.61667	22.85	21										
13	2993 V. cundina	6.155	-75.3736	Colombia	Antioquia	13	13	-75.3736	6.155	17.1875	9.208333	91.17162	29.08647	22.2	12.1	10.1	17.01667	17.05	21										
14	3339 V. cundina	6.5166	-75.5	Colombia	Antioquia	14	14	-75.5	6.5166	17.00833	10.2	85	54.01319	22.9	10.9	12	16.75	16.63333											
15	3350 V. cundina	5.9833	-75.35	Colombia	Antioquia	15	15	-75.35	5.9833	14.90417	8.325	91.48362	22.30556	19.5	10.4	9.1	14.78333	14.75	16										
16	3055 V. cundina	5.7256	-73.7472	Colombia	Boyaca	16	16	-73.7472	5.7256	13.45	9.75	81.93277	35.291	19.4	7.5	11.9	13.98333	13.66667											
17	3461 V. cundina	5.7408	-73.7375	Colombia	Boyaca	17	17	-73.7375	5.7408	13.45	9.75	81.93277	35.291	19.4	7.5	11.9	13.98333	13.66667											

- To construct graphs of the two-dimensional climatic niches, reorganizing the columns according to the following is recommended: Column A: *ID*; Column B: *SPECIES*; Column C: the climatic variable for the X-axis of the graph (in this case: *BIO1, Annual mean temperature*); Column D: the climatic variable for the Y-axis of the graph (in this case: *BIO12, Annual precipitation*).

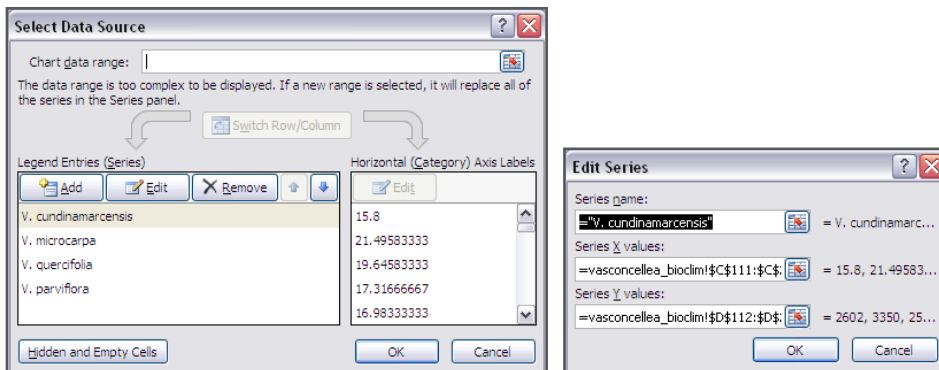
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
ID	SPECIES	bio1	bio12	LATITUDE	LONGITUDE	COUNTRY	ADM1	RecNo	PointNo	Lon_ext	Lat_ext	bio1	bio2	bio3	bio4	bio5	bio6	bio7	
2	3433 V. cundinamarzensis	15.8	3112	6.9636	-75.4177	Colombia	Antioquia	1	1	-75.4177	6.9636	15.8	10.31667	83.87534	41.99667	21.5	9.2		
3	3037 V. cundinamarzensis	21.49583	2602	7.1716	-75.7633	Colombia	Antioquia	2	2	-75.7633	7.1716	21.49583	10.25833	83.40108	40.81323	27.7	15.4		
4	2816 V. cundinamarzensis	19.64583	3390	7.0333	-75.3166	Colombia	Antioquia	3	3	-75.3166	7.0333	19.64583	9.775	78.83065	45.45019	25.2	12.8		
5	2836 V. cundinamarzensis	16.98333	2277	6.9	-75.9666	Colombia	Antioquia	4	4	-75.25	6.3333	17.31667	9.566667	90.25157	34.06767	22.5	11.9		
6	3030 V. cundinamarzensis	18.0625	2015	6.1666	-75.5866	Colombia	Antioquia	5	5	-75.9666	6.9	16.98333	10.2	85	41.79677	22.9	10.9		
7	3418 V. cundinamarzensis	18.0625	2168	6.1666	-75.5866	Colombia	Antioquia	6	6	-75.5866	6.1666	18.0625	9.625	89.95327	34.78015	23.6	12.9		
8	3336 V. cundinamarzensis	17.44167	2168	6.4861	-75.3952	Colombia	Antioquia	7	7	-75.3952	6.4861	17.44167	10.11667	86.46724	51.51493	23.2	11.5		
9	3336 V. cundinamarzensis	17.025	1871	6.2833	-75.4333	Colombia	Antioquia	8	8	-75.4333	6.2833	17.025	9.733333	89.29664	34.21191	22.4	11.5		
10	3027 V. cundinamarzensis	16.39167	1635	6.4627	-75.5563	Colombia	Antioquia	9	9	-75.5563	6.4627	16.39167	10.1	84.87395	50.71459	22.3	10.4		
11	3027 V. cundinamarzensis	20.36667	1636	6.335	-75.5527	Colombia	Antioquia	10	10	-75.5527	6.335	20.36667	10.98333	88.57527	39.808	26.7	14.3		
12	3027 V. cundinamarzensis	15.25833	2024	6.8477	-75.4608	Colombia	Antioquia	11	11	-75.4608	6.8477	15.25833	10.48633	86.73611	42.983	21.1	9.1		
13	3354 V. cundinamarzensis	22.97917	2230	5.8	-75.5666	Colombia	Antioquia	12	12	-75.5666	5.8	22.97917	10.675	87.5	45.79889	29.5	17.3		
14	2993 V. cundinamarzensis	17.1875	2271	6.155	-75.3736	Colombia	Antioquia	13	13	-75.3736	6.155	17.1875	9.208333	91.17162	29.08647	22.2	12.1		
15	3339 V. cundinamarzensis	17.00833	1808	6.5166	-75.5	Colombia	Antioquia	14	14	-75.5	6.5166	17.00833	10.2	85	54.01319	22.9	10.9		
16	3350 V. cundinamarzensis	14.90417	2389	5.9833	-75.35	Colombia	Antioquia	15	15	-75.35	5.9833	14.90417	8.325	91.48362	22.30556	19.5	10.4		
17	3055 V. cundinamarzensis	13.45	1436	5.7256	-73.7472	Colombia	Boyaca	16	16	-73.7472	5.7256	13.45	9.75	81.93277	35.291	19.4	7.5		
18	3461 V. cundinamarzensis	13.45	1436	5.7408	-73.7375	Colombia	Boyaca	17	17	-73.7375	5.7408	13.45	9.75	81.93277	35.291	19.4	7.5		

- Using Excel, you can now create a scatter graph representing the different species, the annual mean temperature range in the X-axis and the annual precipitation range in the Y-axis. This allows you to visualize the differences between the species in a two-dimensional niche space. The graph generated in this analysis shows the climate ranges for *V. cundinamarzensis*, *V. microcarpa*, *V. quercifolia* and *V. parviflora*.

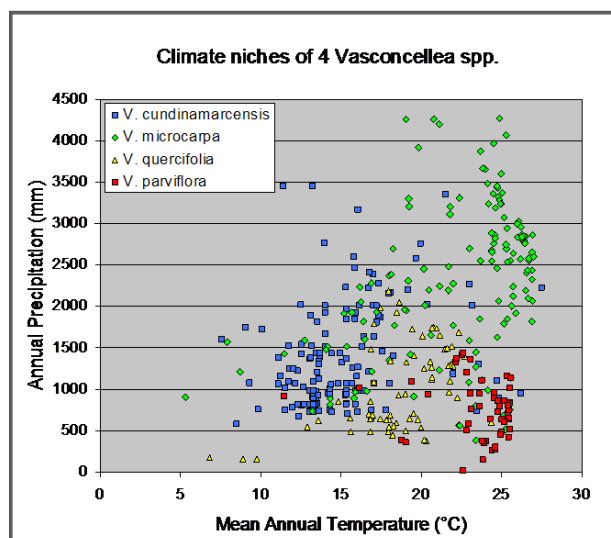
24. Select the temperature values (*X values*) and precipitation values (*Y values*) for each species as a separate series.
 - a. Selection of series in Excel 2003.



- b. Selection of series in Excel 2007.



Final result:



The graph representing the two-dimensional niches for the different *Vasconcellea* species should be similar to the graph above. Differences in annual precipitation and mean annual temperature in the realized niches are clearly observed. The *V. cundinamarcensis* and *V. microcarpa* niches are large in comparison to the *V. quercifolia* and *V. parviflora* niches, suggesting that these species have adapted to a wider range of environments. The *V. quercifolia* realized niche is limited to temperate zones with moderate annual rainfall (this is a species typical of the Interandean valleys in Bolivia), and the *V. parviflora* niche is limited to hot and drier environments (coastal areas in northern Peru and Ecuador), suggesting the species is well adapted to locations with high levels of environmental stress.

Similarly, a multivariate niche with all 19 climatic variables can be used to describe differences between the species. Multivariate analyses such as the Principal Component Analysis (PCA) are beyond the scope of this manual but do provide additional information which may not be possible to observe when using only two dimensions. The relevance of using DIVA-GIS is that the data of all 19 climatic variables can be extracted from the locations of presence points, after which such data can be further analyzed using statistical programmes and software such as Genstat (<http://www.vsnr.co.uk/software/genstat>) or R (<http://www.r-project.org>).

6.2. Modelling the potential distribution of a species

Often, the available set of presence data does not cover the entire range of a species' natural distribution. Species distribution modelling programmes such as Maxent (Phillips et al. 2006) enable one to approximate the full distribution range. These programmes are practical tools to identify those areas where a species is likely to occur. The results of the species distribution modelling analysis can be used for different combined spatial analyses, e.g. evaluating the impact of climate change on the distribution of species (which will be discussed in the next section), identifying collection areas (explained in Section 6.4), or identifying suitable zones for crop and tree production (as mentioned in the introduction of this chapter).

Species distribution modelling programmes identify sites with similar environments to those where a species has already been observed as potential occurrence areas. The data required to identify these potential distribution areas include species presence points as well as the rasters of environmental variables covering the study area. First, a niche is defined based on the environmental values that correspond to the presence points used in the analysis. Then, the similarities between the environmental values at a specific cell and those of the niche of the modelled species are calculated for each raster cell in the study area. With this information, the model calculates the probability of a species' occurrence in each raster cell. Even though the next analysis is based on climate data, Maxent and other niche programmes also allow one to include other types of variables in the model (e.g. soil variables).

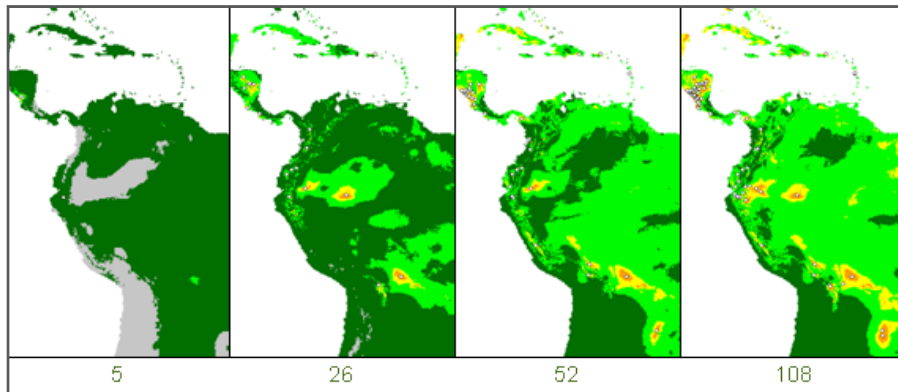
As mentioned, the *Modeling* menu of DIVA-GIS includes the *Predict* option (see Section 6.1) which provides for two integrated species distribution modelling programmes: *Bioclim* and *Domain*; these are different from Maxent. Each programme uses different statistical methods to estimate the realized niche and to calculate the probability of species' occurrence in each raster cell. Therefore, results are likely to differ. Maxent calculates the species' realized niche and probability of occurrence using an algorithm for maximum entropy (Phillips et al. 2006). As Maxent has fared well in evaluations, in comparison to other programmes (Elith et al. 2006; Hernandez et al. 2006), it is the programme of choice

in this manual to undertake species distribution modelling analyses.

It is important to realize that when a geographical area shows environmental conditions favourable for a species, this does not necessarily mean that the species actually occurs in this area. Dispersal limitations because of the species' reproduction system and geophysical barriers can prevent a species from occupying all the geographic areas showing an environment similar to that of its realized niche. It is also true that a species may not be present in areas where it could possibly occur, if its natural habitat has been altered by human interference.

The importance of presence point data quantity

As mentioned in the introduction of Section 6.1, a sufficient number of presence points is crucial for obtaining sound modeling results. The illustration above shows an example of potential distribution maps of *Carica papaya* generated in DIVA-GIS using the *Bioclim* modeling tool. Results of potential distribution are established as the number of points increases in the modeling process (from 5 to 108 presence points). After 50 points, prediction of potential distribution stabilizes and does not change significantly, even if more presence points are included.



PROGRAMMES AND DATA FILES TO USE IN THIS SECTION

Programmes:

- DIVA-GIS
- Excel
- Maxent and Java

Data Files:

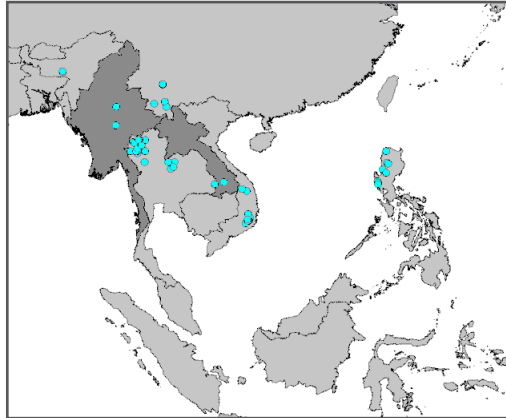
- Folder 6.2 Potential distribution
- *pkesiya.csv*
- *seacountries (shp, shx, dbf)*
- Folder *wclim_sea_2-5min (asc files)*

6.2.1. How to model the potential natural distribution of a plant species

Pinus kesiya is a dominant species of natural pine forests in Southeast Asia and is of economic importance. Observations in many countries suggest a broad distribution of this species; however, the available observations are dispersed. For some countries, only one or two presence points are available. Species distribution modelling enables the identification of the potential range where this species may occur naturally (van Zonneveld et al. 2009a). This analysis will model the potential distribution of *P. kesiya*. You will learn how to utilize Maxent to model the potential natural distribution of a species and to visualize the results generated by Maxent in DIVA-GIS.

The observed distribution of *P. kesiya*

The map below shows the available presence points for *P. kesiya*. In some countries, such as Laos and Myanmar (in dark gray) presence points for this species are scarce (one and two presence points, respectively, in this dataset). Maxent applies a predictive model, based on these distribution points, to identify the areas where this species could potentially occur. After completing this analysis, it is apparent that, despite limited data points, extensive areas within these countries have a high probability of *P. kesiya* occurrence.



Using Maxent to model potential natural distribution of a species

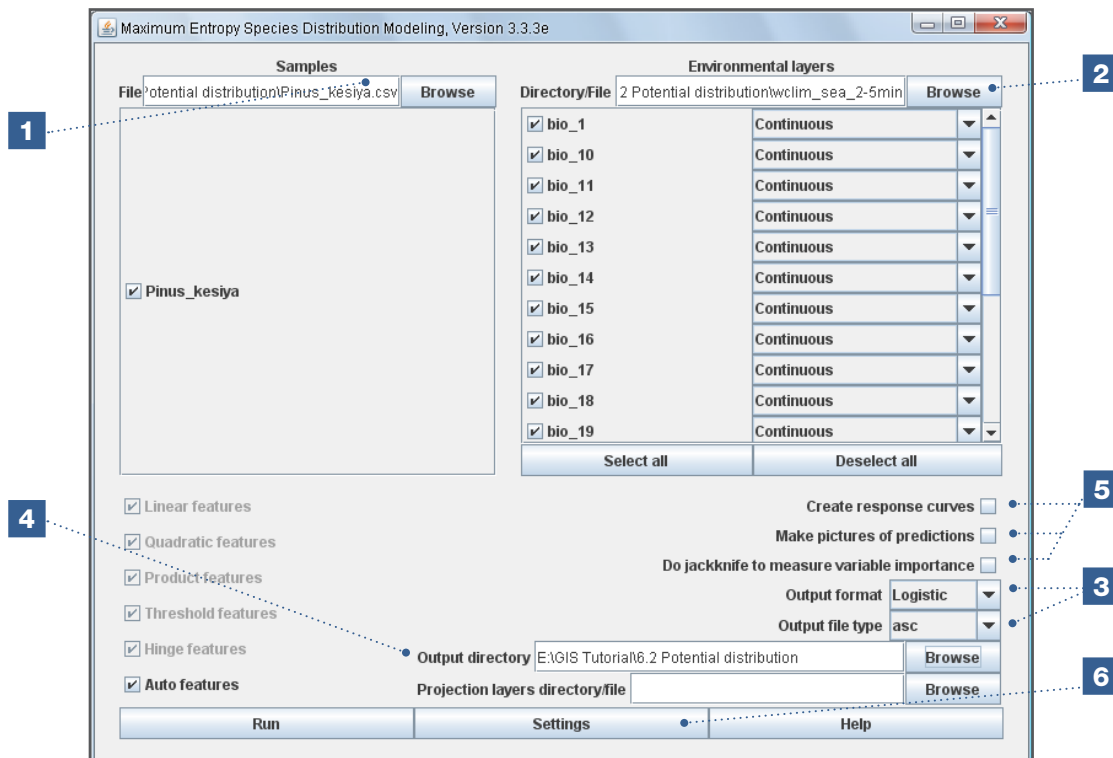
Steps:

1. In Maxent, enter the *pkesiya.csv* file with the presence points in the *Samples* window.

Note

Many files will be generated when performing this analysis. Carefully name and save these files so they can be easily located as they will be used frequently in this section.

2. In Maxent, select under *Environmental layers* the folder *wclim_sea_2-5min* to import the climate layers in ASCII format (*.asc). In this analysis, the rasters have a 2.5-minute resolution which represents the value of the 19 Bioclim variables in the study area (Southeast Asia). See section 2.2.2 for further details about importing climate data in Maxent.
3. Maxent generates a raster of potential distribution. Different output file types can be selected (under: *Output file type*), but the use of the ASCII file type (*.asc) is recommended.
4. In the *Output directory* window, select the location (file path) where the results of the modelling will be saved.
5. We recommend unchecking the option *Make pictures of predictions* and maintain all other default options to assure that Maxent runs well. For further information on these options, please refer to the Maxent manual (Philips, 2009).
6. Go to *Settings* tab to modify the parameters.



Note

In order for Maxent to process the environmental rasters, all rasters must have the same parameters in terms of their properties, resolution and coordinates of corners or vertices. If this is not the case, Maxent will generate an error and will not be able to run.

Maxent only uses the information in the first three columns (see 2.1.3). If more columns (*Fields*) are included in the CSV file (*.csv), the *Visual warnings* sign will be displayed automatically. In this case, check *OK* in each field. Alternatively, select the *Suppress similar visual warnings* option.

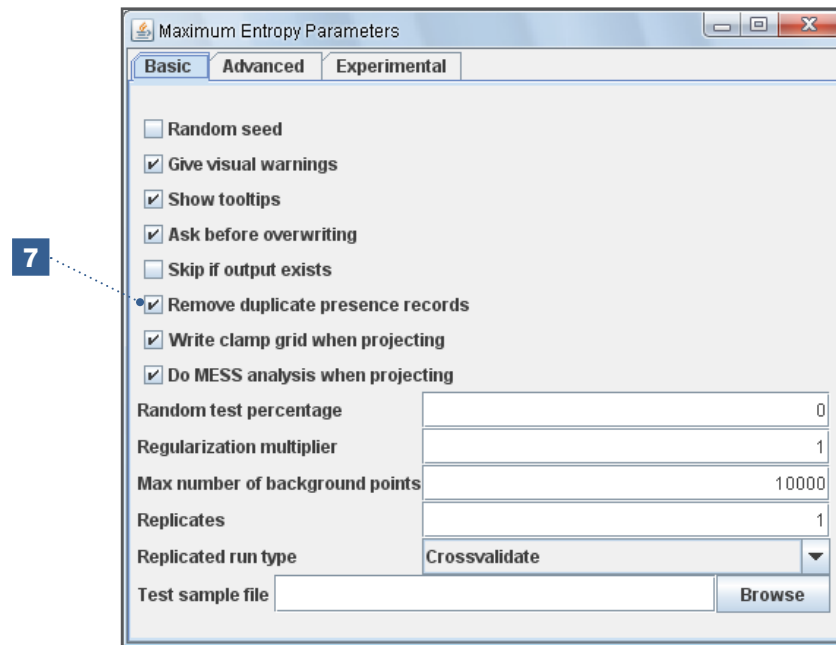
As mentioned in Step 3, Maxent also allows one to select the raster output files in *.grd format. The advantage of using *.grd files, as compared to raster files in ASCII format (*.asc), is that these can be opened directly in DIVA-GIS. However, errors have sometimes been observed in Maxent when using (*.grd) output files. Therefore, using ASCII output files is recommended in this manual.

Maxent will read all raster files in the folder selected as the *Directory File* under *Environmental layers*. When additional raster files (*.grd), are stored in the *Directory File* folder they are automatically included in the list of environmental layers. This can generate errors or undesired results. This may occur if ASCII files are created in DIVA-GIS based on *.grd files (see Chapter 3) and stored in the same folder as the original raster files (*.grd). By saving the relevant ASCII files for the analysis in a separate folder or de-selecting all undesired raster files in the list of environmental layers in Maxent, such errors can be avoided.

Maxent is capable of conducting an analysis for several species at once when presence points for each species are saved in the same CSV file (*.csv). This is explained in Section 6.4.

The *Settings* option allows you to modify the conditions under which Maxent generates a potential distribution model. A relevant parameter for this analysis is the *Remove duplicate presence records* option, found under the *Basic* tab. For further information about the *Settings*, please refer to the Maxent manual (Phillips 2009).

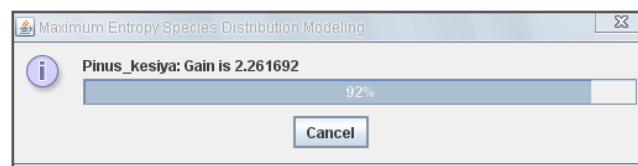
7. With the *Remove duplicate presence records* option, duplicate presence points in one raster cell are removed from the analysis to reduce sampling bias, which would favour the climatic conditions of those sites where sampling was highly concentrated. For this analysis, make sure that this option is selected (checked) and that the other basic settings correspond to what is shown in the screenshot below.



8. After modifying the *Settings* option, return to the main window. Click *Run* to start calculating the species' potential area.

Running Maxent

When you press the *Run* button, a progress monitor appears which describes the steps being taken. After the environmental layers are loaded and the initialization process is complete, progress towards developing the model will be displayed:

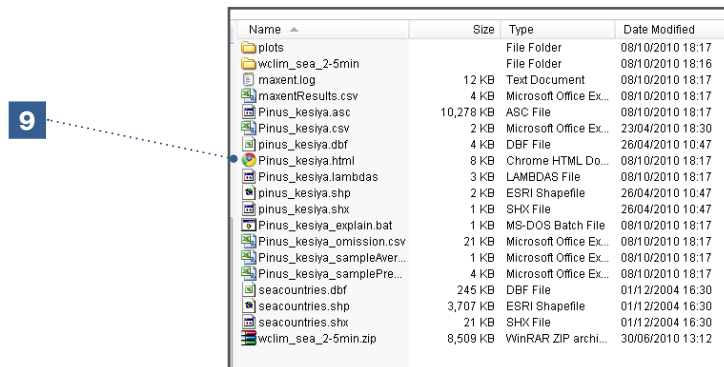


The gain is closely related to deviance, a measure of goodness of fit. The higher the gain, the more discriminative the predicted distribution for species' occurrence is in comparison to a random distribution. For example, if the gain is two (2), it indicates that the average likelihood of the presence samples is $\exp(2) \approx 7.4$ times higher than that of a random background pixel. For further information refer to the Maxent background paper (Phillips et al. 2006) and the Maxent manual (Phillips 2009).

Results from Maxent

Results are saved in the folder selected under the *Output Directory*. One of the saved files is an HTML document, which summarizes all results. The analysis of the most important parameters is briefly described below. For more information on the analysis, see Anderson (2003) and Phillips et al. (2006).

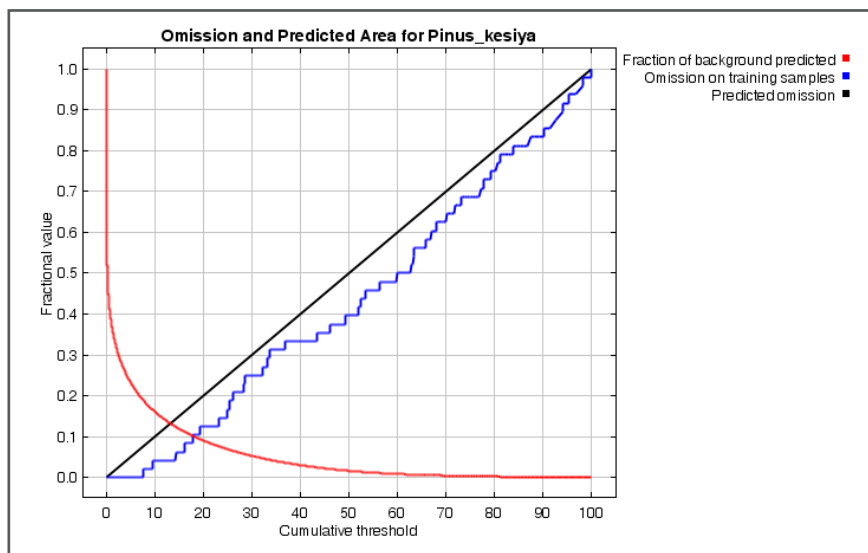
9. Open the HTML document: *Pinus_kesiya* in your internet browser (e.g. Firefox, Chrome, or Internet Explorer).



10. The omission rate is a statistic indicating model performance. The *Omission and Predicted Area* plot consists of three lines:

- *Omission on training samples* (blue line) shows the fractions of the presence points located outside the potential area as modelled by Maxent from low to high threshold values limiting the predicted area of occurrence (*Cumulative threshold*). Training samples is synonymous to presence points.
- *Fraction of background predicted* (red line) shows the fractions of background points from the study area included in the modelled distribution area under varying *Cumulative thresholds*.
- *Predicted omission* (black line) is a reference line.

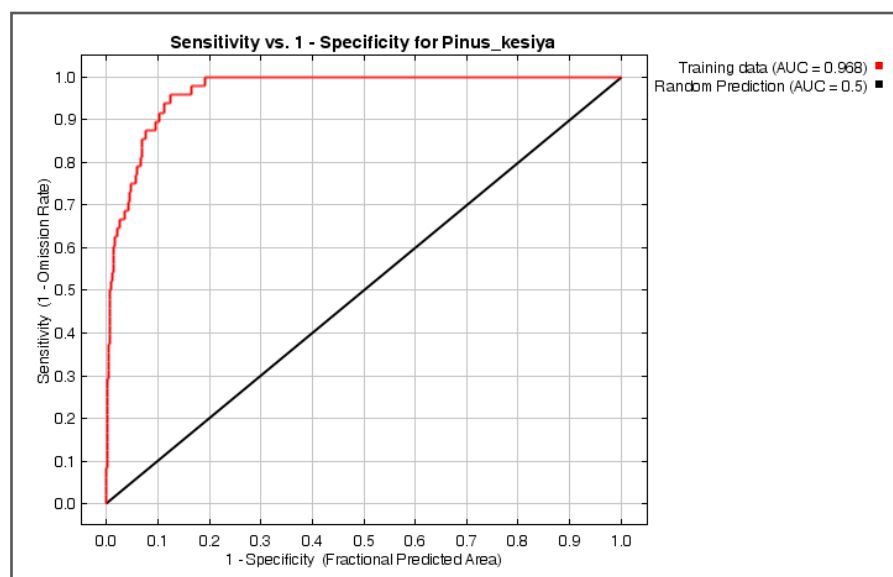
If the blue line (*Omission on training samples*) is well below the black line (*Predicted omission*), this might indicate some over-fitting because of dependence between presence points.



11. One of the parameters used for evaluating the predictive ability of the models generated by Maxent is the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The AUC is equal to the likelihood that a randomly selected presence point is located in a raster cell with a higher probability value for species occurrence than a randomly selected absence point. Species distribution modelling in this manual includes presence points only. To still be able to calculate the AUC, Maxent replaces absence points with randomly generated points from the study area. The AUC can then be interpreted as the likelihood that a randomly selected presence point is located in a raster cell with a higher probability value for species occurrence than a randomly generated point (Phillips et al. 2006).

The fractional predicted area on the X-axis of the AUC graph is the fraction of the total study area where the species is predicted present, while the sensitivity on the Y-axis is the proportion of presence points in the modeled area of occurrence on the total number of actual presence points (Phillips 2009).

The highest predictive power of a model generated by Maxent is reached when the AUC has a value of 1. In practice, no AUC will be lower than 0.5, which is similar to *Random prediction*. In that case Maxent has no predictive power at all. Araújo et al. (2005) recommend the following interpretation of AUC for the models generated: Excellent if $AUC > 0.90$; Good if $0.80 > AUC \leq 0.90$; Acceptable if $0.70 > AUC \leq 0.80$; Bad if $0.60 > AUC \leq 0.70$; Invalid if $0.50 > AUC \leq 0.60$. In the case of this analysis, AUC is 0.963. If the predicted area is low in comparison to the study area, high AUC values doesn't necessarily reflect good model performance, and simply could be an artifact of the AUC statistic (Phillips 2009). For information about AUC and the ROC curve, please refer to Fawcett (2006).



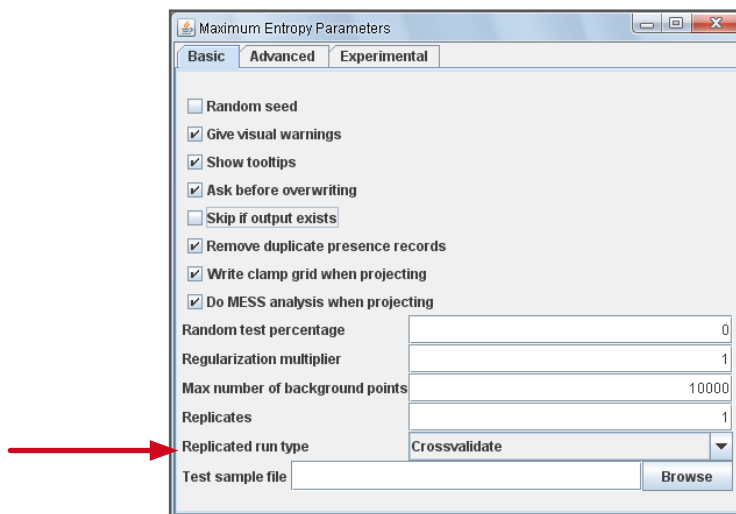
The table in the next step (Step 12) illustrates how thresholds can be used to limit the potential area of a species (in this case, *Pinus kesiya*). These thresholds represent the minimum probability of a species potentially occurring in the environment of a specific cell. This concept assumes that sites with a probability above the threshold have climatic conditions appropriate for the occurrence of the species, while the species would not occur in sites below the threshold. The higher the threshold, the more restricted the potential distribution areas. There is no standard threshold value and the user must define this parameter [for additional information on thresholds for potential distributions and niche limits, see Lui et al. (2005)].

12. In this analysis, the *10 percentile training presence* threshold will be used: the probability value at which 90 % of the presence points fall within the potential area. The remaining 10 %, which fall outside the potential area, are those with an atypical environment, not included within the limits of the realized niche. Depending on the *Output format* selected (see Step 5), select either the *Cumulative* or the *Logistic threshold* (default output format is *Logistic threshold*). For this analysis, select the *Logistic threshold* (see red arrow).

Cumulative threshold	Logistic threshold	Description	Fractional predicted area	Training omission rate
1.000	0.013	Fixed cumulative value 1	0.377	0.000
5.000	0.049	Fixed cumulative value 5	0.232	0.000
10.000	0.091	Fixed cumulative value 10	0.160	0.042
7.483	0.071	Minimum training presence	0.190	0.000
17.715	0.154	10 percentile training presence	0.102	0.083
17.715	0.154	Equal training sensitivity and specificity	0.102	0.104
14.266	0.128	Maximum training sensitivity plus specificity	0.124	0.042
3.994	0.040	Balance training omission, predicted area and threshold value	0.254	0.000
17.403	0.152	Equate entropy of thresholded and original distributions	0.104	0.083

Validation of the model's robustness

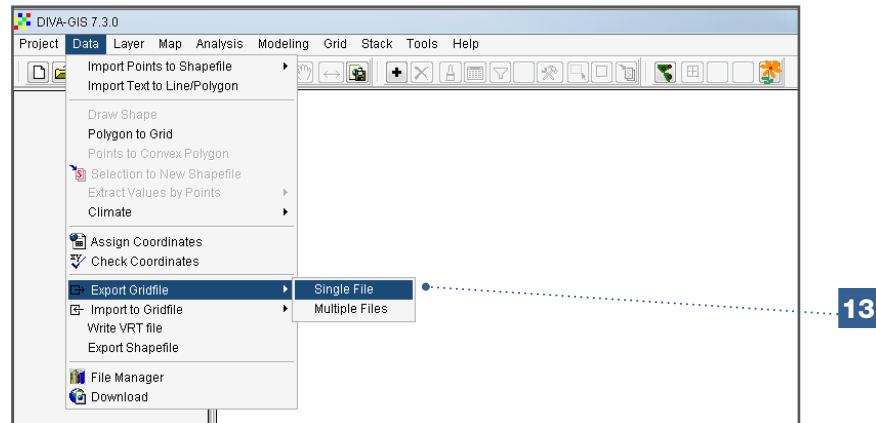
The robustness of the model developed by Maxent can be validated using one of the methods available under the option *Replicate run type* in the *Basic* tab. [Validation of models is not dealt with in depth in this manual; for more information consult Araújo et al. (2005) and Philips et al. (2006)]. The robustness or transferability of the model is relevant when predicting potential distribution areas outside the observed distribution and when using different climate scenarios, such as predictions of species distribution under future climates.



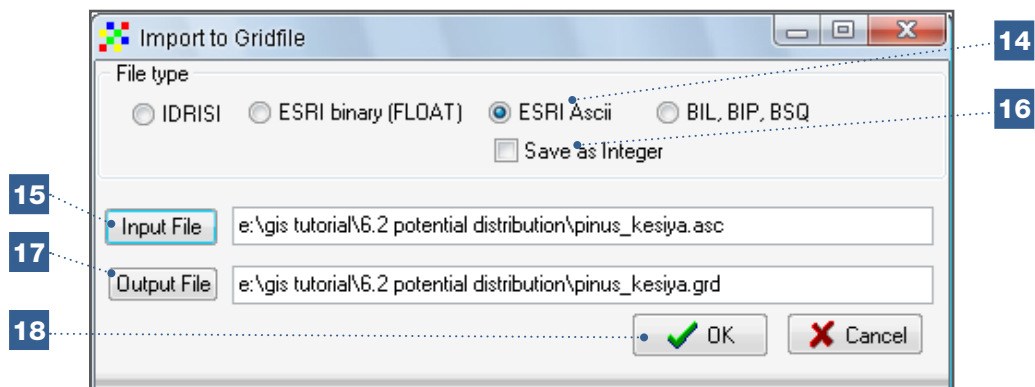
Visualizing the results of Maxent in DIVA-GIS

The raster of the potential distribution of *Pinus kesiya* generated by Maxent is in ASCII format (*.asc). It can be found in the output folder (the same folder where the HTML file is stored). In order to visualize and modify these files, they will be imported to DIVA-GIS.

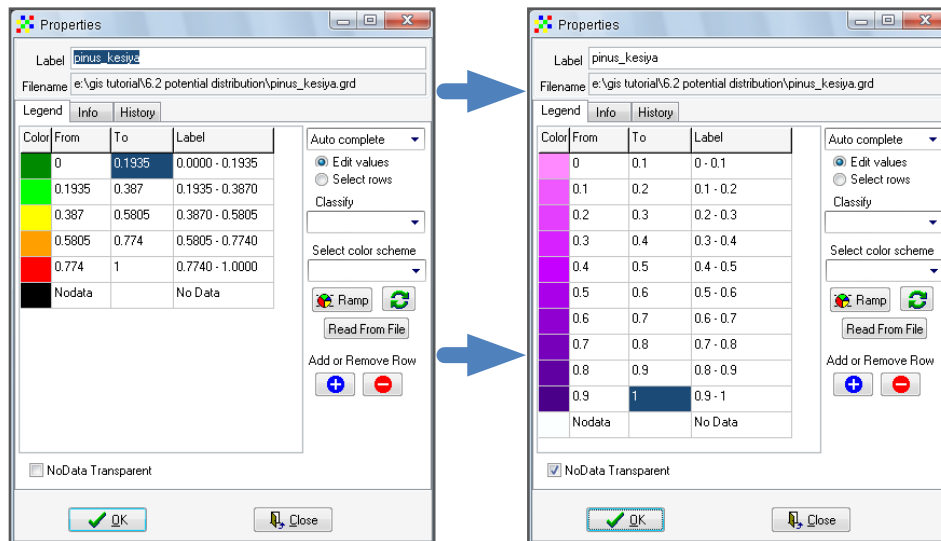
- Open the *Import to Gridfile/Single File* option in the *Data* menu to import the rasters in ASCII format (*.asc) to DIVA-GIS (see Chapter 3).



- In the *File type* window, select *ESRI ASCII*.
- Select the desired ASCII raster file (*.asc) from which to generate a raster file in *.grd format. For this analysis, select the ASCII file generated by Maxent: *Pinus_kesiya.asc*.
- Under *Save as Integer* you can select whether or not to generate a raster in which values are presented in integers in the legend of DIVA-GIS. For this analysis, the raster will not have integral values; therefore, it is important *not* to select this box.
- Under *Output File*, give the name for the raster file (*.grd) that will be generated by DIVA-GIS.
- Click *OK* to initiate the process.

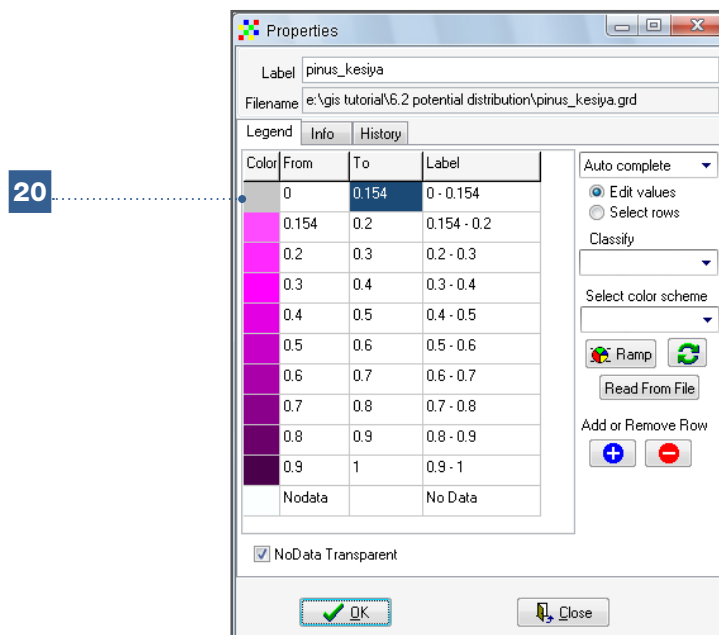


19. After the raster generated in Maxent has been imported to DIVA-GIS, this file can be opened, visualized and modified. Visualization can be improved by modifying the standard legend, as described in Chapter 3. The following steps outline the type of changes which can be made; the choice of colours and the use of a more gradual scale can vary according to the user's preferences.



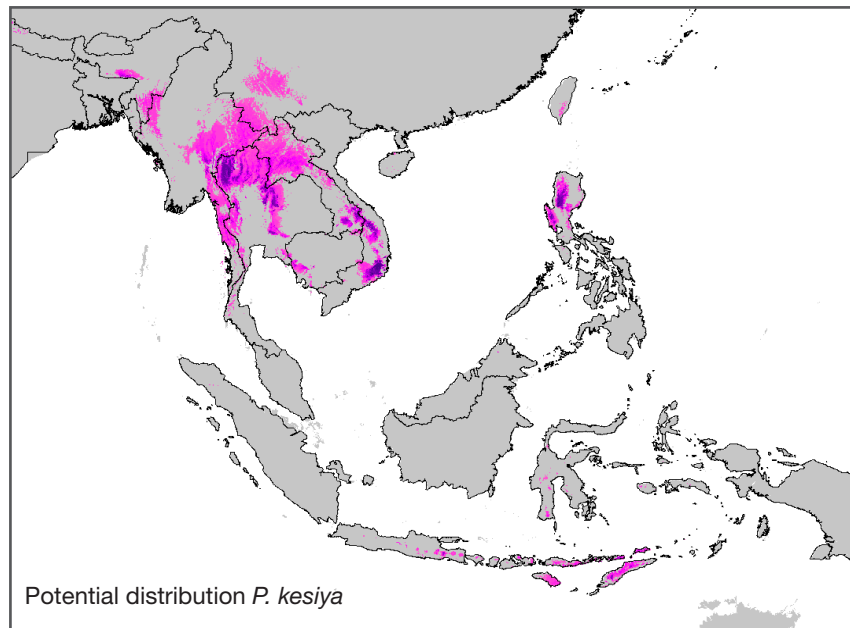
Visualizing the threshold limiting potential distribution areas

20. The 10 percentile training presence threshold, found in the table of thresholds generated by Maxent, will be used in this analysis (as explained in Step 11). The threshold value for *P. kesiya* is 0.154, as per the table of thresholds. Go to the legend of the raster and create a new class ranging from zero (0) to the threshold value; select a neutral colour for this new class.

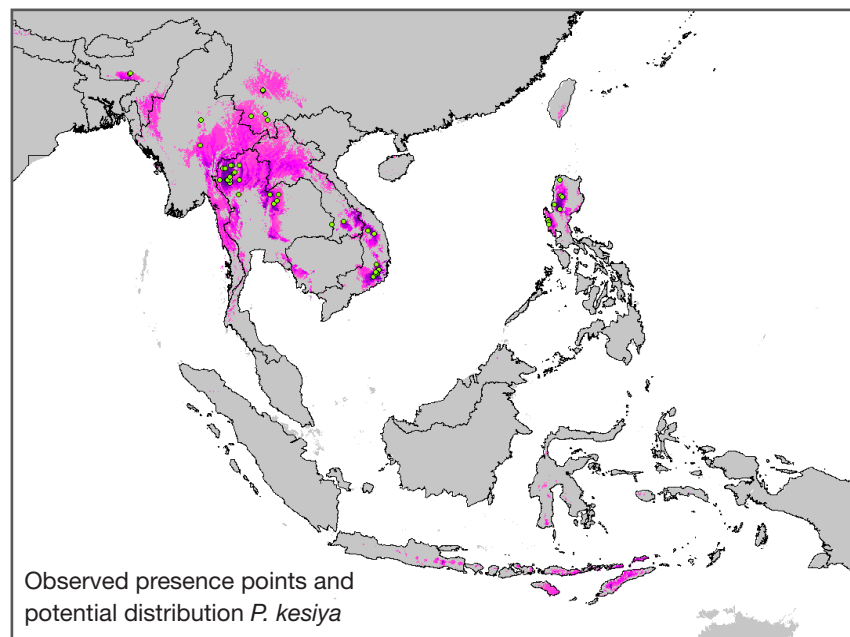


21. An administrative unit layer (*seacountries.shp*) should be added to facilitate locating the potential distribution areas.

Final result:



After modifying the potential distribution map of *P. kesiya* (as outlined in Chapter 3), it should appear similar to the image above. Some countries, such as the Laos and Myanmar, have few presence points available. Maxent reveals extensive areas in these countries where this pine species may potentially occur. Species distribution modelling conducted with Maxent completes the observed distribution and, in this instance, emphasizes the existing knowledge gaps. This means that the vacant areas require further study in order to determine the location of pine populations and define conservation strategies for this species (this issue will be further discussed in Section 6.4).



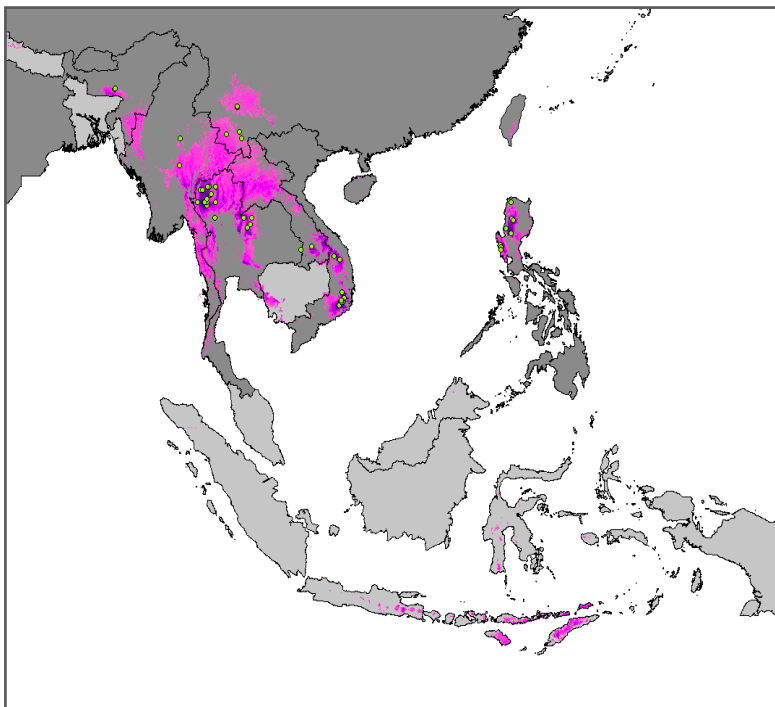
Most likely, the species does not occupy all the potential distribution areas represented by the model because species dispersal is limited by its reproduction system and the presence of geophysical and climatic barriers. In this illustration, the Indonesian

archipelago includes several areas with climatic conditions similar to the realized niche for *P. kesiya*; however, these areas are not included in the species' natural distribution as a result of the limitations noted above. If desired, it is possible to limit potential distribution areas using a fixed-distance buffer area around each presence point [see Willis et al. (2003)]. This approach may adequately reflect the dispersal limitations, due to a species' reproduction system, but will not solve the problem of geophysical and climatic barriers, which can be unexpectedly close to represented presence points. Alternatively, one might consult the literature or contact experts on the species being studied and then compare this information with that provided by the potential distribution model.

Occasionally, the literature will refer to areas where a species occurs naturally, even though the model does not predict its occurrence in these locations. In such instances, the environment corresponding to the database of the presence points may not be completely representative for the climatic niche, and the area is therefore not captured. For example, the *GRIN Taxonomy for Plants* (USDA, ARS, National Genetic Resources Program 2009) reports the presence of *P. kesiya* in Bhutan and the Chinese province of Xizang; however, the model generated by Maxent does not predict the occurrence of the species in these areas. To resolve this discrepancy, a more detailed study and/or contacting local experts is required.

Thus, it should be noted that species distribution modelling, scientific literature and expert data provide complementary information. The combination of these information sources will help to provide a complete picture of the natural distribution of a given species.

Comparison between potential natural distribution of *P. kesiya* according to Maxent and the natural distribution according to literature



The natural distribution of *P. kesiya*, according *GRIN Taxonomy for Plants* (USDA, ARS, National Genetic Resources Program 2009), is defined by administrative boundaries highlighted on the map in dark gray. Countries included in the natural distribution are: China (Xizang, Yunnan), Bhutan, India, Laos, Myanmar, Thailand, Vietnam and the Philippines (Luzon).

6.3. Modelling the impact of climate change on species' distribution

Global climate change is ever more evident (IPCC 2007). Consequently, geographic areas corresponding to biomes, ecosystems and species' ecological niches are changing, which is likely to affect the natural distribution of many species.

Species distribution modelling can be used to provide a rapid evaluation of the potential impact of climate change on the distribution of ecosystems and the species that inhabit them. The process consists of detecting changes in species distribution by comparing the potential distribution areas in the current climate (based on climatic conditions at presence points) with the potential distribution areas, based on a species' current climate preferences, under future climatic conditions. Future potential distribution areas of occurrence are identified using climate layers based on the projections of General Circulation Models (GCM). Climate research institutions from various countries generate these models (<http://www.ipcc-data.org>) which predict future climatic conditions under different emission scenarios developed by the Intergovernmental Panel on Climate Change (IPCC) (for further information see: http://www.ipcc.ch/publications_and_data/publications_and_data.shtml).

Note

Be careful not to develop a future potential distribution map based on the future climate conditions at the (present day) presence points. The current and future potential distribution areas both need to be based on the species climate niche that is calculated with current climate data.

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION

Programmes:

- DIVA-GIS
- Maxent and Java

Data Files:

- Folder 6.3 Climate change
- *pkesiya.csv*
- *seacountries (shp, shx, dbf)*
- Folder *wclim_sea_2-5min (asc files)*
- Folder *gcm_sea_2-5min (asc files)*

6.3.1. How to evaluate the impact of climate change on the distribution of species

Despite several on-going in situ pine conservation projects in Southeast Asia, the area occupied by *Pinus kesiya* has diminished in recent decades. Many remaining stands are continuously threatened by unsustainable resin extraction and wood harvesting practices and are also likely to be threatened by the effects of climate change. Species distribution modelling, along with climate models, can help to determine the most at-risk populations and to identify where the conservation of genetic resources requires urgent measures (van Zonneveld et al. 2009a). In this analysis, the potential impact of climate change on the distribution of *P. kesiya* will be explored.

In this analysis, you will learn how to use Maxent to predict the potential distribution of a species under current and future climatic conditions and to examine the impact of climate change on a species using DIVA-GIS.

This analysis will use the climate projections for the year 2050 under the A2 emission scenario from three different GCMs: *CCCMA*, *HADCM3* and *CSIRO*. Each model has a slightly different projection of the future climate. Therefore, predictions of potential

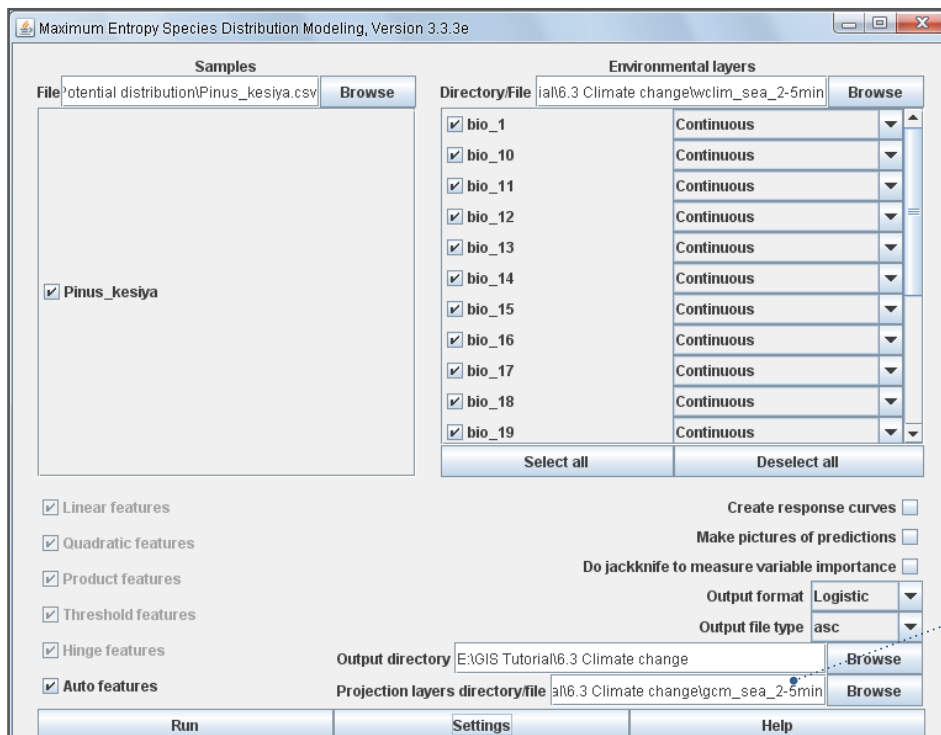
distribution areas for a species will vary depending on the GCM used in Maxent. Here, the average of the three selected GCMs will be used.

Note

Many files will be generated when performing this analysis. Carefully name and save these files for easy access as they will be used frequently in this section.

Steps:

1. To run the Maxent model, follow Steps 1 to 8 outlined in Analysis 6.2.1. Next, in the *Projection layers directory file* window, import the climatic variables under future conditions (*gcm_sea_2-5min*) to the database.

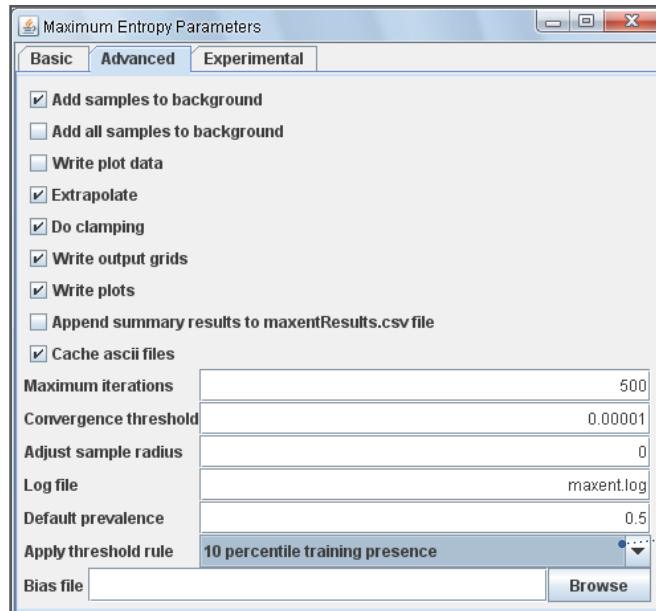
**Note**

In order for Maxent to process the environmental raster files of both current and future conditions (in this case, *wclim_2-5min* and *gcm_sea_2-5min*), all files must have the same parameters in terms of raster properties, resolution and raster corners or vertices.

Apply threshold rule to create binary rasters

Maxent can also generate binary presence (1) and absence (0) rasters of potential distribution areas. This format is useful when layers of potential species' distribution are compared and combined, as will be done in this analysis. The predicted potential distribution areas under current conditions and future projections will be compared, identifying those distribution areas that will be strongly affected by climate change, as well as those areas where the impact will be less severe and new areas for the natural occurrence of the species in the future.

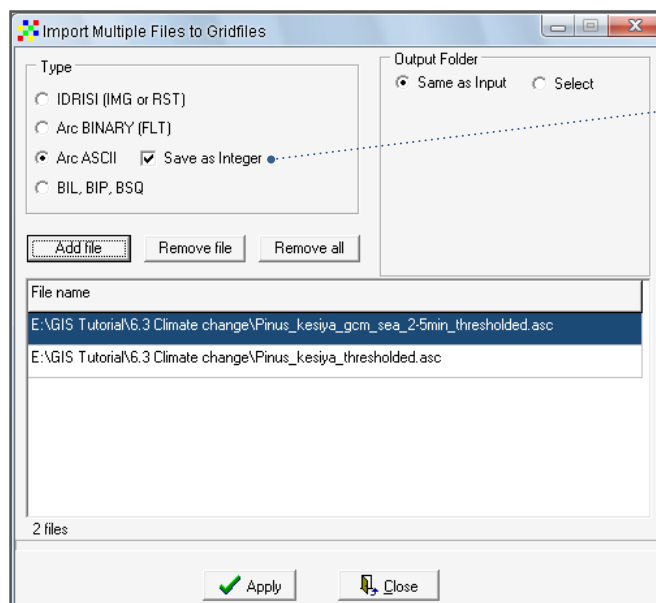
- To generate binary rasters in Maxent, go to the *Advanced settings* menu and, under the *Applied threshold rule*, select a threshold that limits the potential area. This threshold (*10 percentile training presence*) is the same as that which was manually visualized using DIVA-GIS in Step 21 of Analysis 6.2.1. Run the analysis in Maxent.



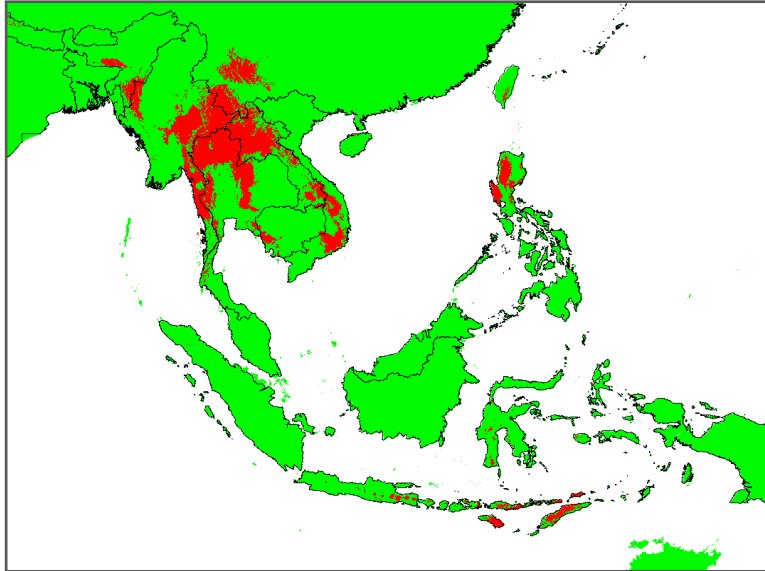
Examine the impact of climate change using DIVA-GIS

In the first part of this section, binary rasters under the current climate and future projection were generated using Maxent in ASCII format (*.asc). These rasters will now be imported to DIVA-GIS as was done in Steps 13 to 19 of Analysis 6.2.1.

- Open DIVA-GIS and import the presence/absence rasters of the potential distribution areas (*Pinus_kesiya_gcm_sea_2-5min_thresholded.asc* and *Pinus_kesiya_thresholded.asc*) as integer value raster files (these files can be found in the output folder). Check the *Save as Integer* option under *Import to Gridfile/Multiple Files*.



4. After closing the *Import to Gridfile/Multiple Files* window, open the rasters using DIVA-GIS. The presence/absence map generated for *P. kesiya* under current climate conditions (*Pinus_kesiya_thresholded.grd*) should be similar to the following illustration.



Overlaying rasters of current and future potential distribution areas

A useful way to identify the impact of climate change on the distribution of groups of species is to overlay the rasters of current and future potential distribution areas. Binary rasters are recommended for constructing maps which are easy to interpret; however, rasters that gradually show potential distribution areas with increasing probabilities can also be used for more detailed analysis. In this analysis, the binary rasters of current and future potential distribution areas of *P. kesiya* will be overlaid.

Overlaying binary rasters results in four possible situations for each cell:

- i. *High impact areas*: areas where a species potentially occurs in the present climate but which will not be suitable anymore in the future.
- ii. *Areas outside of the realized niche*: areas that are neither suitable under current conditions nor under future conditions (as modelled).
- iii. *Low impact areas*: areas where the species can potentially occur in both present and future climates.
- iv. *New suitable areas*: areas where a species could potentially occur in the future, but which are not suitable for natural occurrence under current conditions.

Each binary raster has two values: presence (1) and absence (0). When these two values are added or subtracted, the only possible results for the cells are: negative one (-1), zero (0) and one (1). However, a problem exists in that a fourth value is required to represent the four possible situations outlined above. The following table illustrates this problem: subtracting the rasters results in cells with the same value for the second and third situations (ii - areas outside of the realized niche and iii - low impact areas), as neither experiences any change [hence the value of zero (0) when combining].

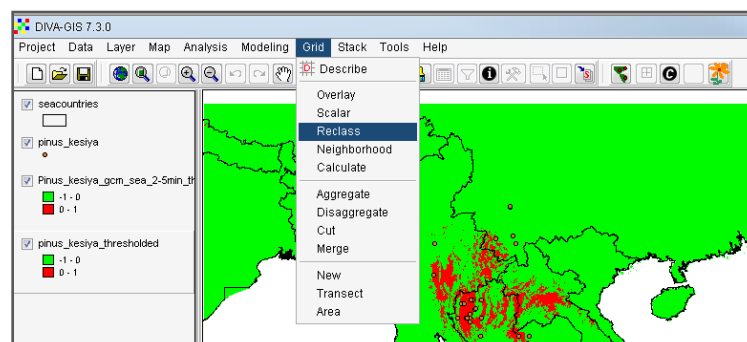
Situation	Raster of future potential distribution areas (cell value)	Raster of current potential distribution areas (cell value)	Result after subtracting rasters (cell value)
(i) High impact areas	0	1	-1
(ii) Outside of realized niche	0	0	0
(iii) Low impact areas	1	1	0
(iv) New suitable areas	1	0	1

To solve this problem in DIVA-GIS, go to the *Reclass* option and change the cell value for potential distribution areas from one (1) to two (2) for one of the two rasters (see Analysis 3.1.3). The following table reflects how this change in cell value will result in a different cell value for all four situations.

Situation	Raster of future potential distribution areas (cell value)	Raster of current potential distribution areas (cell value)	Result after subtracting rasters (cell value)
(i) High impact areas	0	1	-1
(ii) Outside of realized niche	0	0	0
(iii) Low impact areas	2	1	1
(iv) New suitable areas	2	0	2

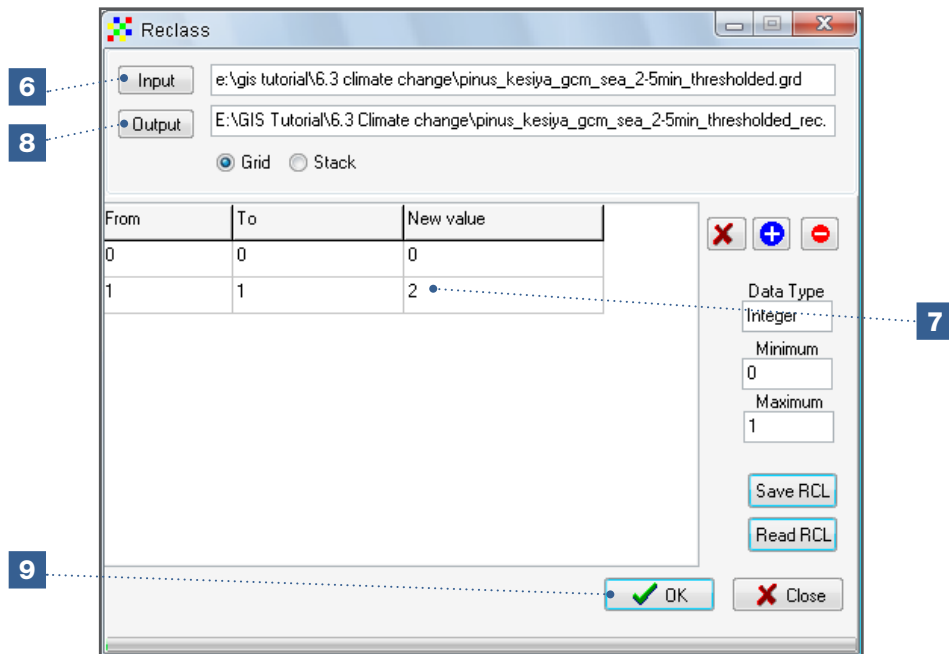
For the next analysis, change the value for the potential distribution areas from one (1) to two (2) for the rasters of the predicted future distribution of *Pinus kesiya*.

- Go to *Grid/Reclass* to change the raster cell value of potential distribution areas under future climate (*.Pinus_kesiya_gcm_sea_2-5min_thresholded.grd) from one (1) to two (2).

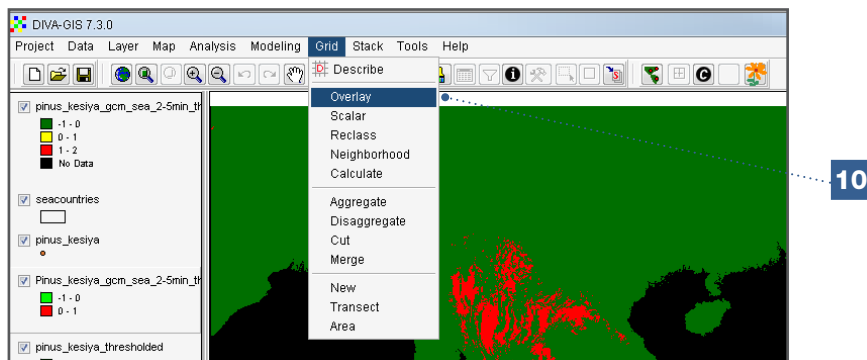


- Under *Input*, select the raster for which you will change the cell values.
- To avoid confusion during the reclassification, change the respective cell values in the *From* column from negative one (-1) to zero (0) and zero (0) to one (1). Then change the cell values from one (1) to two (2) under the *New value* column.

8. Indicate where the raster with the new cell values will be saved. Save the raster under a different name. Normally, this is done by adding the suffix (*rec*).
9. Save the raster with the new cell values.

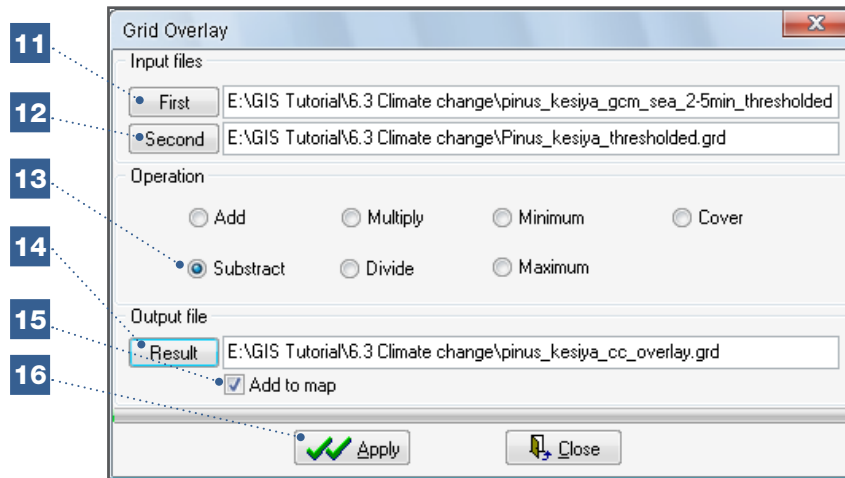


10. After having changed the values, go to *Grid/Overlay* and superimpose the rasters for current and future potential distribution areas.

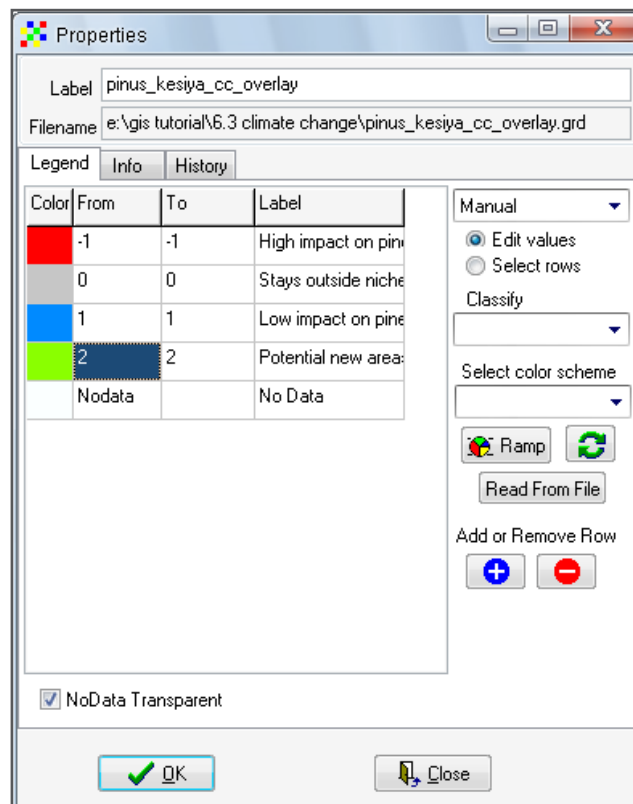


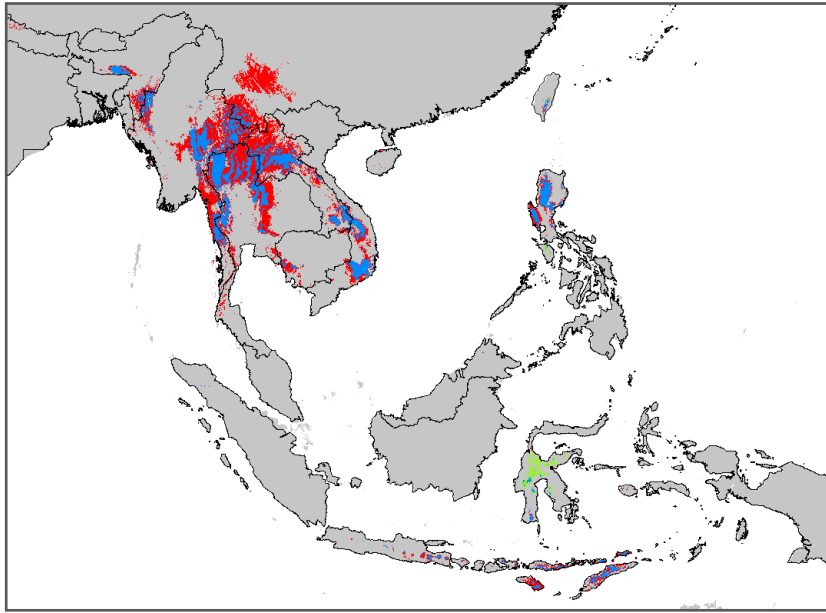
11. Select the binary raster for future potential distribution areas under the *First* tab.
12. Select the binary raster for current potential distribution areas under the *Second* tab.
13. Under *Operation*, select the *Subtract* option to subtract the values of the two rasters.
14. Under the *Result* tab, indicate the raster containing the results from subtracting the overlaid rasters.
15. Select the *Add to map* option. The new raster will then be automatically opened in DIVA-GIS after clicking *Apply*.

16. Click *Apply* to start calculations.



17. After the raster has been opened in DIVA-GIS, the legend and labels can be modified. To easily visualize the four situations, it is recommended to modify the legend as illustrated in the following image.





After editing the map (as explained in Chapter 3), the result should be similar to the map above. The results reveal that climate change will particularly affect *P. kesiya* populations in the Chinese province of Yunnan (due to an expected increase in seasonality) and will impact areas at lower altitudes more generally (as these are predicted to become too hot).

As such, these areas may be prioritized when developing conservation strategies, including those designed to collect germplasm to ensure the *ex situ* conservation of genetic resources before existing stands disappear under the prevailing changes in climate. *In situ* conservation strategies to protect populations predicted to be highly affected might focus on improving the connectivity between fragmented populations to ensure gene flow of adaptive genes. An alternative option would be to assist the migration of these species to newly suitable areas (shaded in green on the map above). In the case of *P. kesiya*, areas of the Indonesian island of Sulawesi and the Philippine island of Mindoro are expected to be suitable for *Pinus kesiya* occurrence by 2050 as a result of climate change. An additional *in situ* conservation strategy might be to increase efforts to conserve populations in low impact areas, where models predict the species will survive in 2050, protecting them from threats caused by human interference.

These predictions only deal with potential impact. The models may overestimate the impact of climate change as species may possess the capacity to adapt to a range of climatic conditions. Several tree species have a high degree of genetic variability and may be able to tolerate a broad range of climates. Multi-site trials conducted with pine species such as *P. kesiya* have shown that the species adapts well to a broad range of climates and is also likely to adapt to new climatic conditions, even though studies conducted using species distribution modelling predict future conditions to be inadequate (van Zonneveld et al. 2009b). It should be noted that soil conditions, competition, predators and other factors also influence the presence of a species and represent additional limitations to the species' current distribution and possible future displacements. However, since climate is considered to be the main driving force affecting distribution areas in the future, models predicting the effects of climate change have not yet focused on or included these other factors.

In spite of their limitations, envelope models are considered to be a useful tool in establishing an initial appreciation of the potential impact of climate change on the distribution of species (Pearson and Dawson 2003).

6.4. Identification of gaps in collections of wild plant species

As mentioned in Section 6.2, one of the uses of potential distribution maps is to detect gaps in the data on a species' distribution. A gap refers to a location where species distribution modelling predicts that a species could potentially occur, but where specimens and/or germplasm of wild species or crops have not actually been collected. Gaps may indicate that accessions and specimens from these areas are missing in germplasm or herbarium collections (e.g. Jarvis et al. 2005; Scheldeman et al. 2007). On the other hand, local studies and observations of the species in these areas may be available, but the information may not be broadly disseminated and not included in initiatives such as the Global Biodiversity Information Facility (GBIF) (www.gbif.org) that promote the use of data by the general public (see Section 2.3).

It is a real possibility, however, that a species simply does not exist in the area predicted, due to dispersal limitations. Such a case was explored in Section 6.2 where *Pinus kesiya* was predicted to be present in several south-eastern islands of Indonesia but did not naturally occur in those zones.

Some species may also have disappeared from an area due to deforestation, selective extraction or other anthropogenic pressures on the natural habitat. For example, *Polylepis* forests in Bolivia and Peru are currently very fragmented; based on ecological niche studies, these forests were demonstrated to have previously been extensively distributed in the Andean zone of these countries (Fjelds  2002).

PROGRAMMES AND DATA FILES TO USE IN THIS SECTION	
<p>Programmes:</p> <ul style="list-style-type: none"> • DIVA-GIS • Maxent and Java 	<p>Data Files:</p> <p>Folder 6.4 Gap analysis</p> <ul style="list-style-type: none"> • <i>Vasconcellea.csv</i> • <i>Vasconcellea species (shp, shx, dbf)</i> • Folder wclim_ams_5min (asc files)

6.4.1. How to identify possible gaps in collections

Areas of observed diversity of *Vasconcellea* species in Latin America were identified in Section 5.1. By using species distribution modelling, in addition to this observed diversity, a map of the potential diversity of these species can be generated. Areas where it is likely to encounter a diversity of *Vasconcellea* species, but where there are currently few or no records of observations, can be identified by comparing maps of observed and potential diversity. These gaps are areas of particular interest for germplasm collection missions (Scheldeman et al. 2007). The next analysis illustrates how to use DIVA-GIS to create a map of potential diversity and to compare existing gaps between a species' observed and potential diversity.

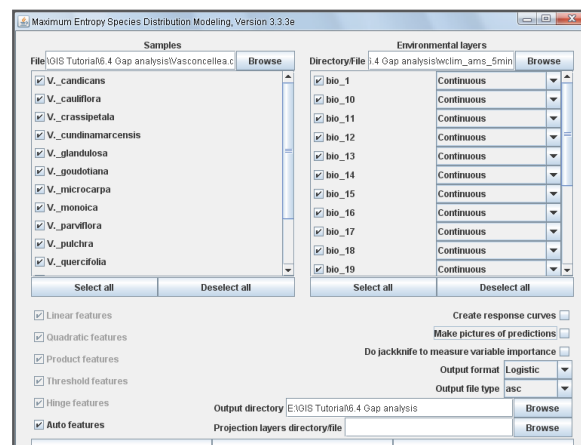
In the previous sections, rasters of potential distribution areas for the natural occurrence of a single species were explored. Rasters of potential diversity can also be generated based on the realized niches of several species by using a stack of binary rasters with potential distribution areas for each individual species.

Note

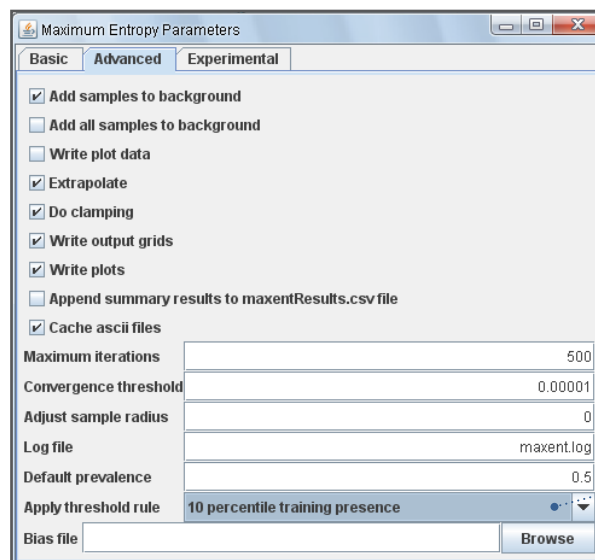
In order for Maxent to process the environmental raster files, all files must have the same parameters in terms of raster properties, resolution and raster corners or vertices.

Steps:

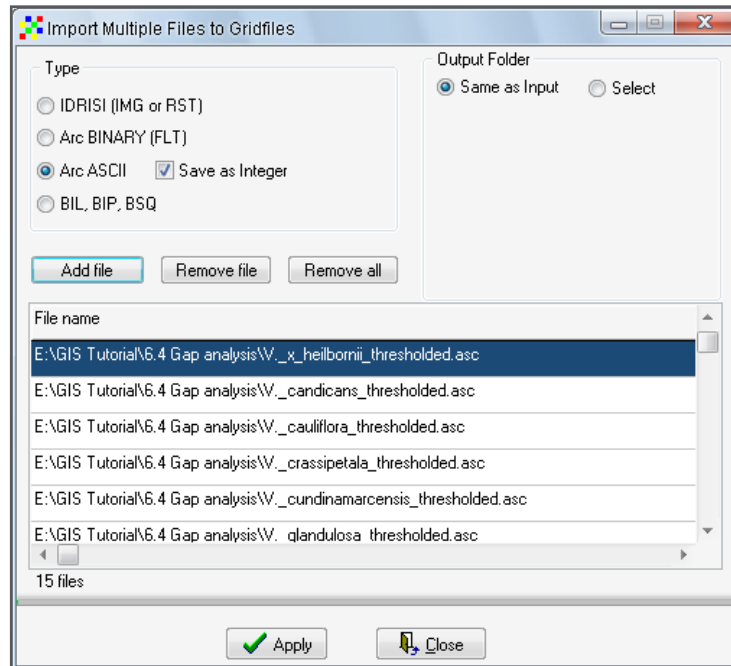
1. Use Maxent to generate potential natural distribution models for all the different *Vasconcellea* species, following the explanation given in Steps 1 to 8 of Analysis 6.2.1, where the potential distribution for one species was modelled. Maxent can also carry out an analysis for multiple species simultaneously when presence points for each species are saved in the same CSV file (*.csv). This is the case for the 15 *Vasconcellea* species (*Vasconcellea.csv* file). For this analysis, models will be generated for the *Vasconcellea* species using the rasters of the Bioclim variables, with a 5-minute resolution, for Latin America and the Caribbean; these can be found in the *wclim_ams_5min* folder.



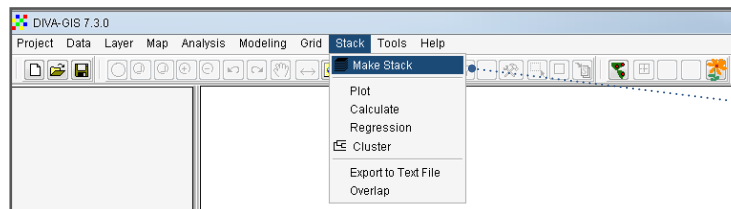
2. To generate binary rasters of potential distribution areas using Maxent, a similar procedure as was conducted in Step 2 of Analysis 6.3.1 is followed: in the *Advanced settings* window, under the *Apply threshold rule* box, select the *10 percentile training presence* option.



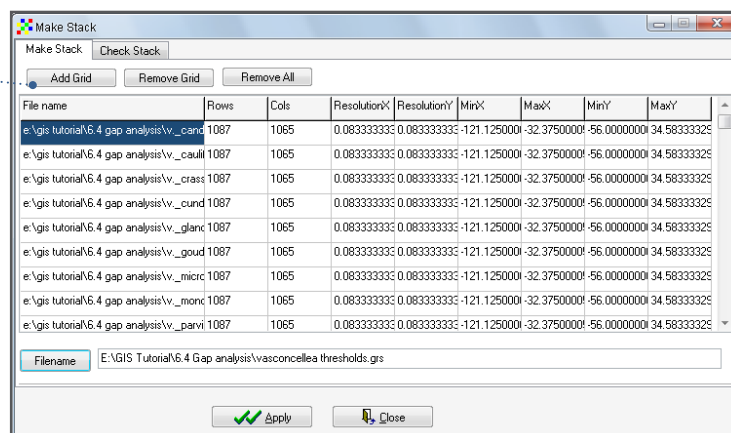
- Import the binary rasters of potential distribution areas to DIVA-GIS, as explained in Steps 13-19 of Analysis 6.2.1 (using: *Import to Gridfile/Multiple Files*).



- Make a stack of the binary rasters (**_thresholded.grd*) by selecting the *Make Stack* option in the *Stack* menu.

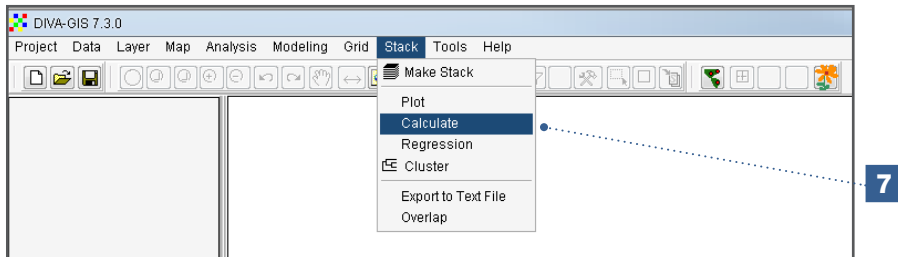


- Indicate the rasters you wish to include in the stack. For this analysis, use the binary raster files of the *Vasconcellea* species (**_thresholded.grd*).

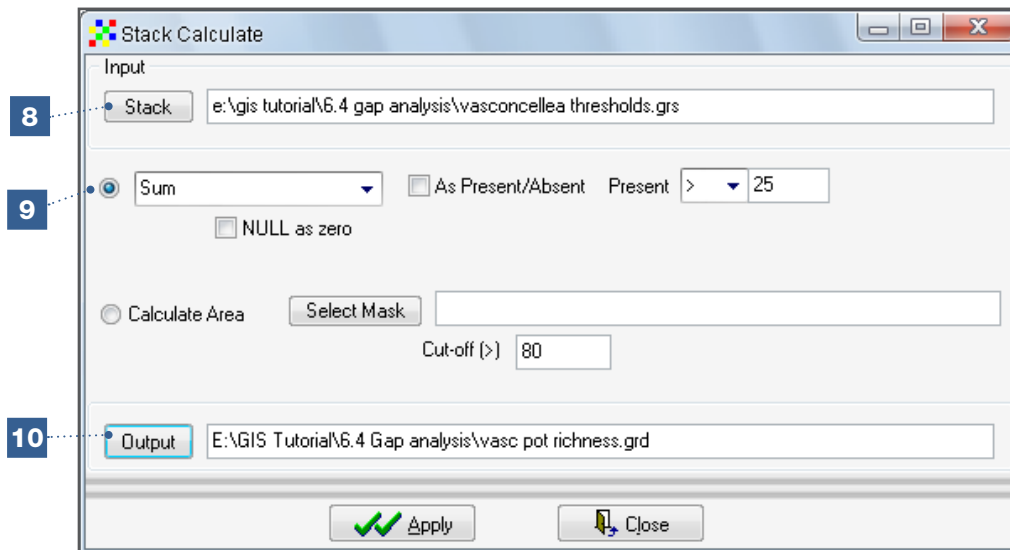


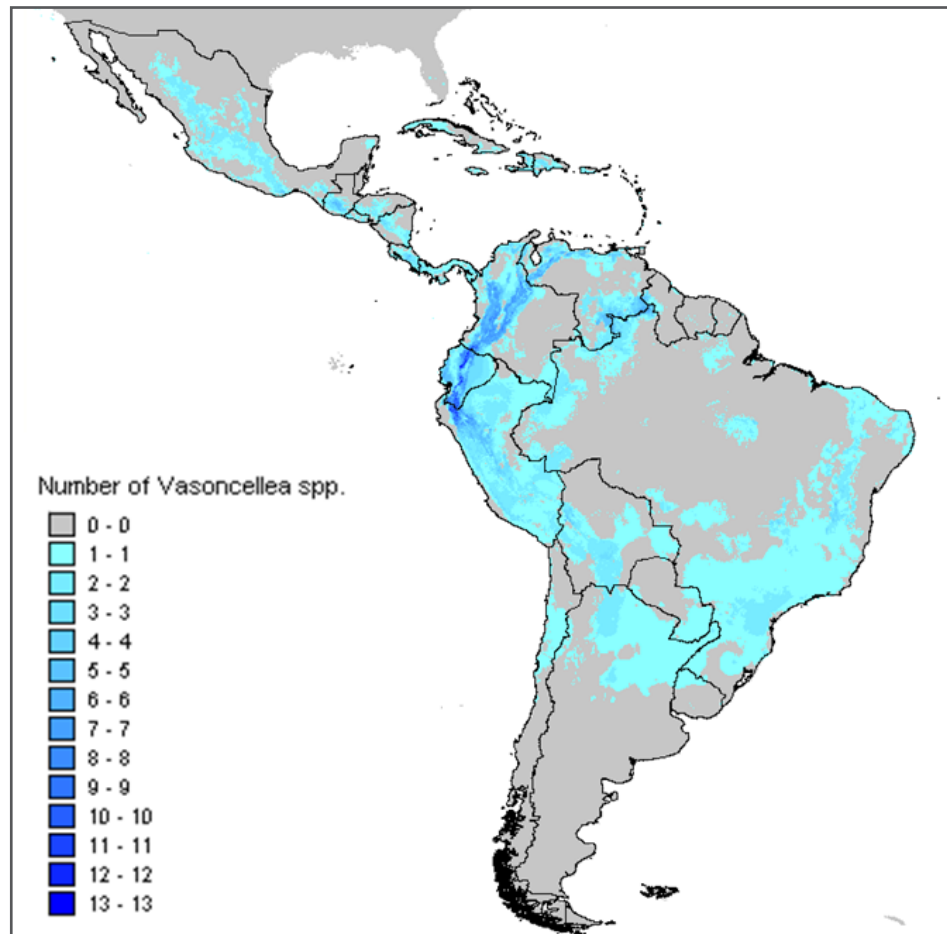
- Assign a name to the stack, click *Apply* and then *Close*.

- To estimate the potential richness of *Vasconcellea* species in Latin America, go to *Stack/Calculate*; this sums the stack of binary rasters representing the potential area of this species.



- Select the stack of rasters for which you wish to make the calculations. In this case we select the stack that was made in step 6.
- Select the *Sum* option to add the binary rasters of the *Vasconcellea* species potential distribution areas.
- Under the *Output* tab indicate the name for the raster file which will be generated.
- Leave all other options as default.
- Click on *Apply* to start the calculating process.





After having edited the map (as explained in Chapter 3), the potential diversity map of the *Vasconcellea* species in Latin America should be similar to the one above. The greatest area for expected diversity is in the northern Andean zone, particularly between Ecuador and Peru and between Ecuador and Colombia.

Gap analysis using DIVA-GIS

To identify gaps in species' distribution, the rasters of potential and observed richness of *Vasconcellea* species are compared. To undertake the gap analysis, you will need:

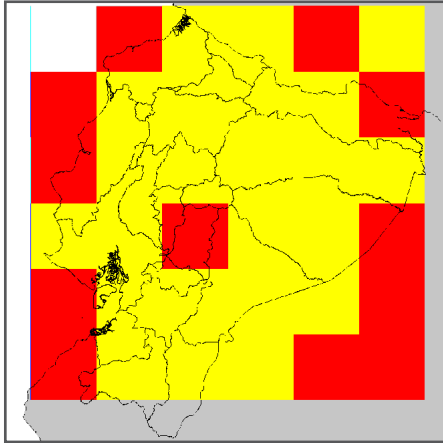
- The raster of potential natural distribution areas for the species.
- An observed distribution raster for the species with the same properties as the raster of potential distribution.

Raster cell size

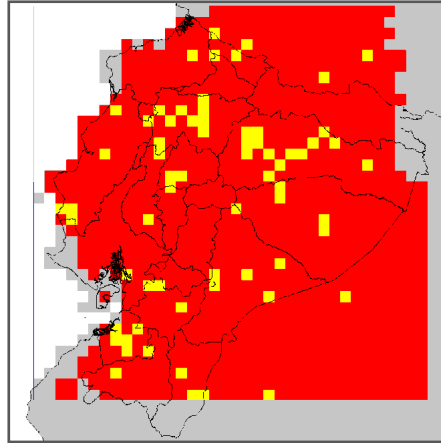
The importance of cell size in detecting spatial patterns has already been discussed in Section 5.1. The cell size is also highly relevant when identifying gaps. A small cell size will serve to detect gaps at a local scale, but when using small cells for a gap analysis in a large study area (e.g. country or regional level), only a limited number of the cells will contain presence points, as most cells will not have been sampled. This will obviously lead to many presumed gaps being identified, complicating the prioritization of sites for additional collection. For studies at a national or regional level a larger cell size is more suitable, yet it will be up to the analyst to decide on the best cell size.

Choice to appropriate cell size in gap analysis

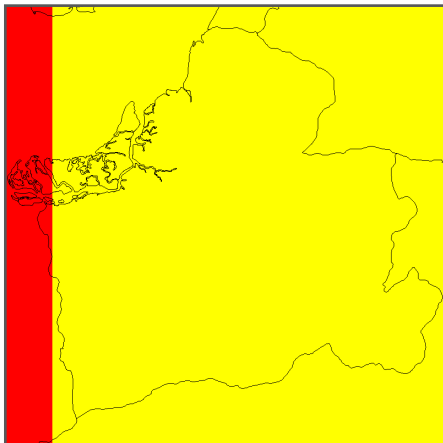
The figures below show a gap analysis for *Vasconcellea microcarpa* in Ecuador using different cell sizes. They illustrate that rasters with a 10-minute cell size can be useful for a gap analysis at local scale (e.g. province) while rasters with a large cell size (1 degree) are more suitable for a gap analysis at national level.



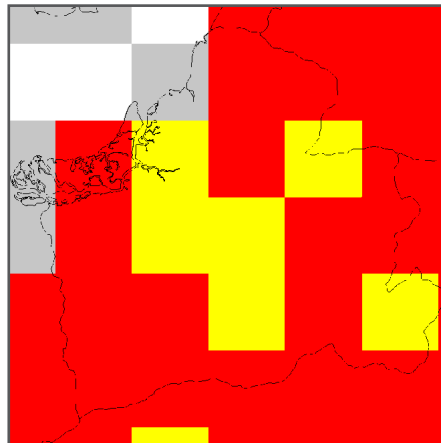
1 degree gap analysis at national level



10 minutes gap analysis at national level



1 degree gap analysis at province level



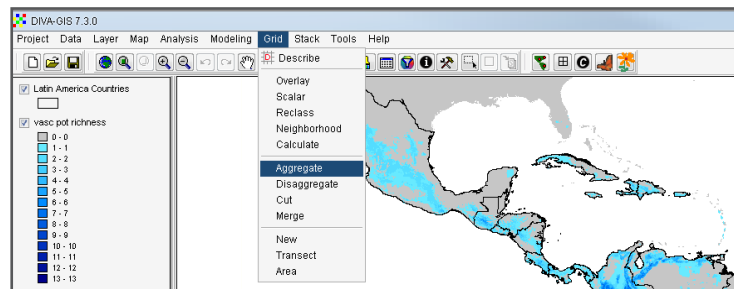
10 minutes gap analysis at province level

- Yellow: areas showing potential presence where the species was actually observed
- Red: potential distribution areas for the species, but where it has not been observed
- Gray: areas outside of the observed and potential presence of the species

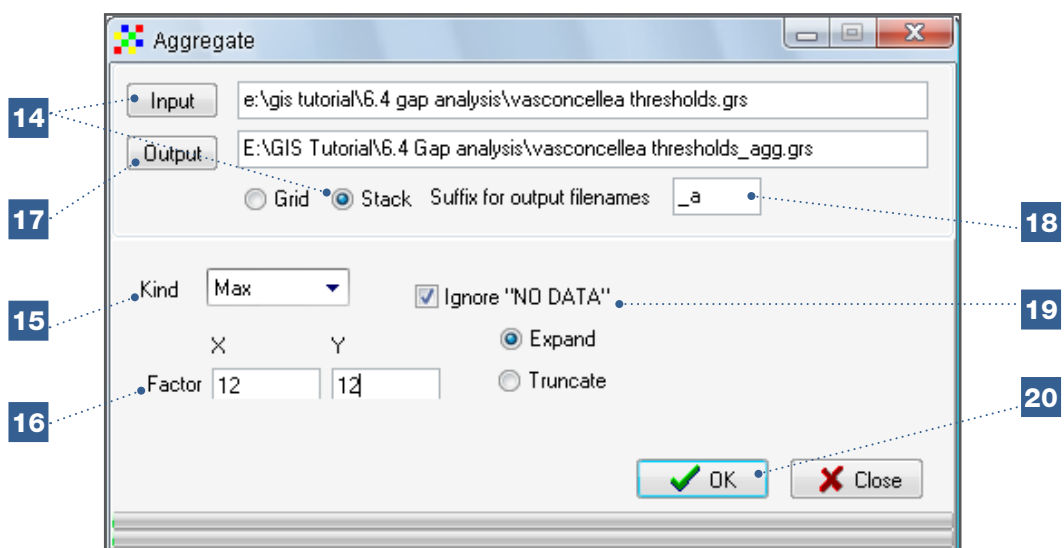
In view of the limited capacity to effectively identify gaps when using high resolution rasters, it is important to increase the cell size of the binary rasters from the previously generated potential *Vasconcellea* diversity raster.

Steps (continued from the previous section):

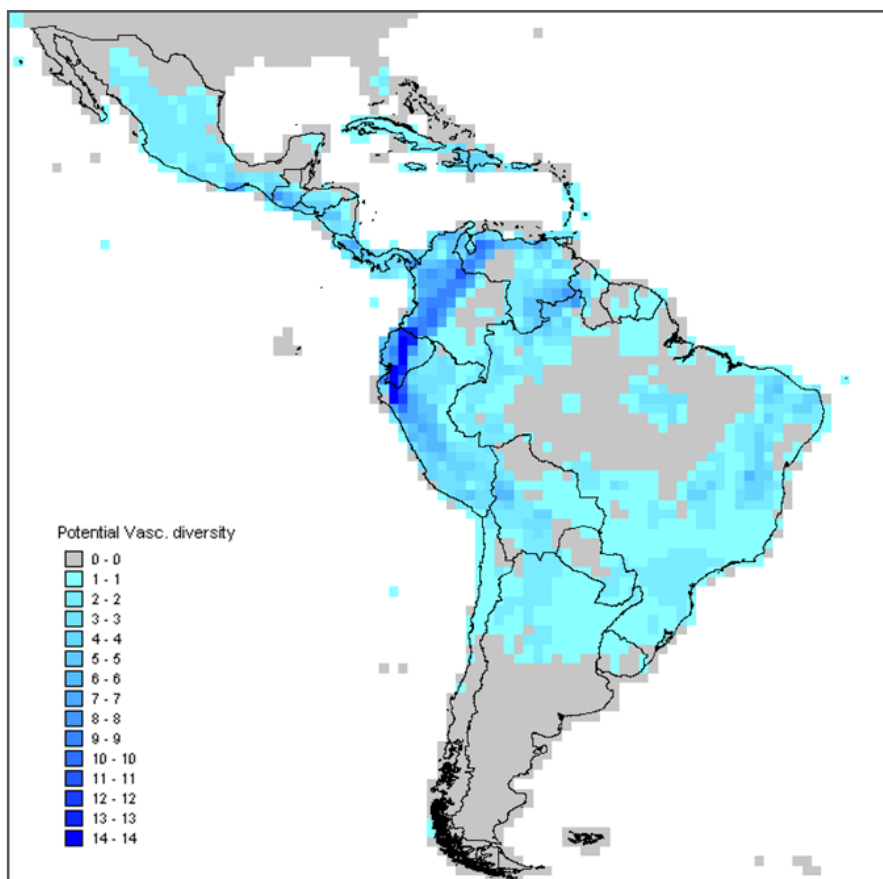
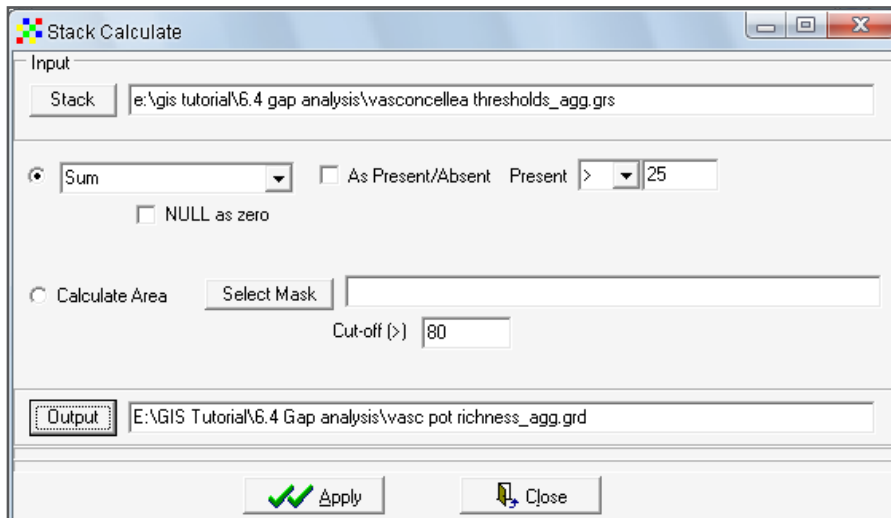
13. In the *Grid* menu, select the *Aggregate* option.



14. In the *Input* box, select the stack of rasters of which to increase the cell size. In this analysis, select the stack with the binary rasters (**_thresholded*) of the potential distribution areas for each *Vasconcellea* species; this is the stack created in Step 4.
15. Under the *Kind* option, select the type of calculation you wish to use to combine the values of the smaller cells when merged to form a larger cell. For this analysis, select the *Max* option so the aggregated cell utilizes the maximum value of its composing cells. If at least one of the composing cells has a value of one (1), the aggregated cell also includes a value of one (1), indicating that the species has been observed in that cell.
16. In the *Factor* box, indicate the extent to which you wish to enlarge the raster cell size. In this example, the 5-minute raster cell resolution will be converted to one (1) degree resolution. To do this, to the cells must be enlarged by a *Factor* of 12 in both directions [since 60 minutes is equal to one (1) degree].
17. Indicate in the *Output* box the file name for the new raster file.
18. You have the option of adding a suffix to the names of aggregated rasters. In this analysis, use the default option: *_a*.
19. Keep the *Expand* and *Ignore "NO DATA"* options selected.
20. Click *OK* to start the calculation process.



21. After increasing the cell size of all rasters within the stack, repeat the process using the *Calculate* option under the *Stack* menu (Steps 7-12 above) to determine how many *Vasconcellea* species are potentially present in each cell.



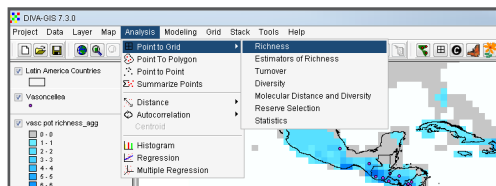
After some modifications and editing, the potential diversity map of the *Vasconcellea* species with the aggregated cells should resemble the one above.

It may seem inefficient to first generate the binary potential distribution rasters in Maxent using detailed climate data (small cell size) only to later aggregate these to a larger cell

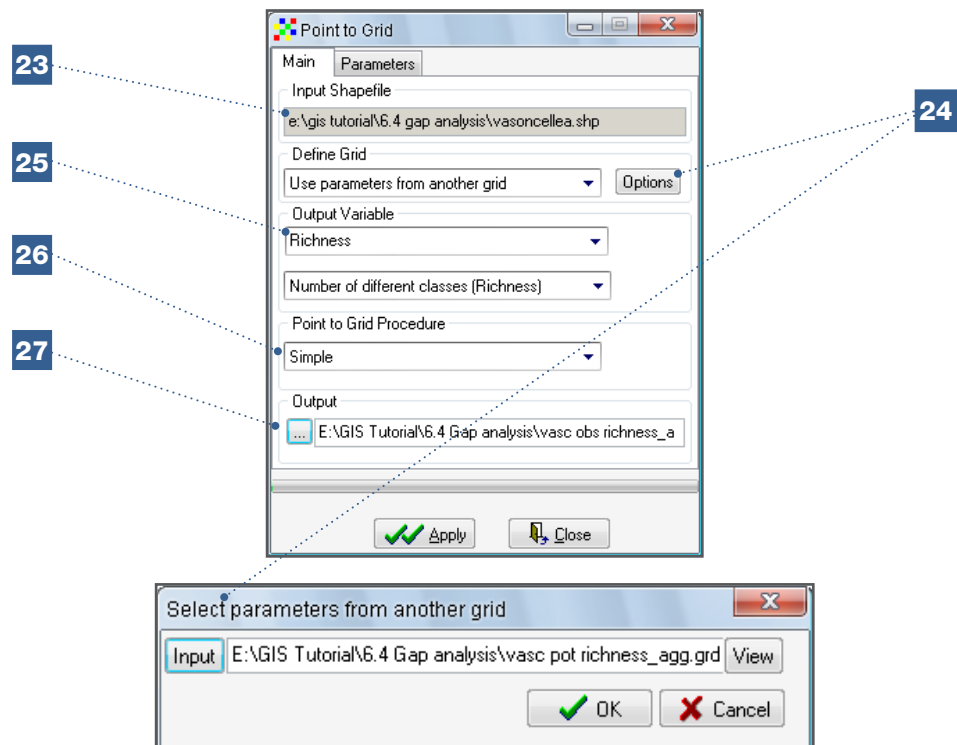
size, rather than immediately generating the binary potential occurrence rasters using larger cells (with climate data of lower resolution). However, it is recommended to use the two-step process described as high resolution climate data usually generates more accurate predictions of potential distribution in Maxent than data of low resolution.

Developing the raster of observed richness of *Vasconcellea* species (see also Section 5.1)

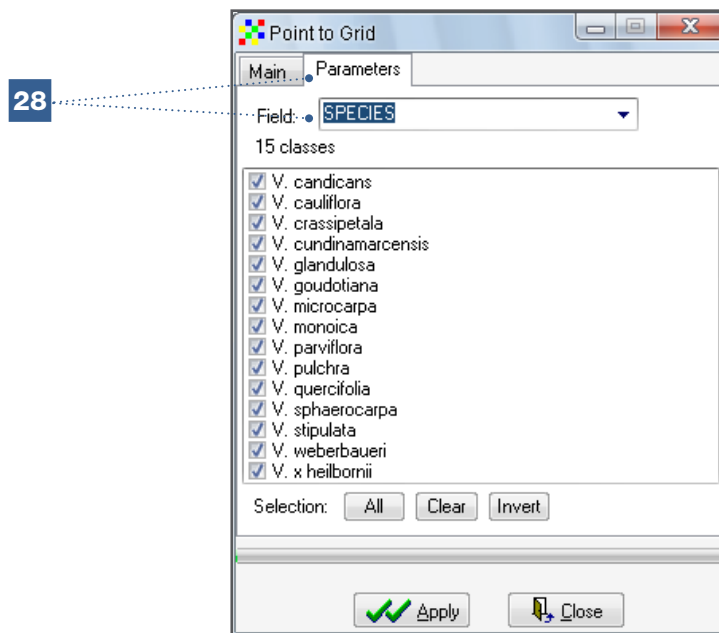
22. Select the vector type file (*.shp) with the presence points of a species or group of species to analyze. In this analysis, the *Vasconcellea* species (*Vasconcellea species.shp*) should be selected.
23. In the *Analysis* menu, go to *Point to Grid/Richness*.



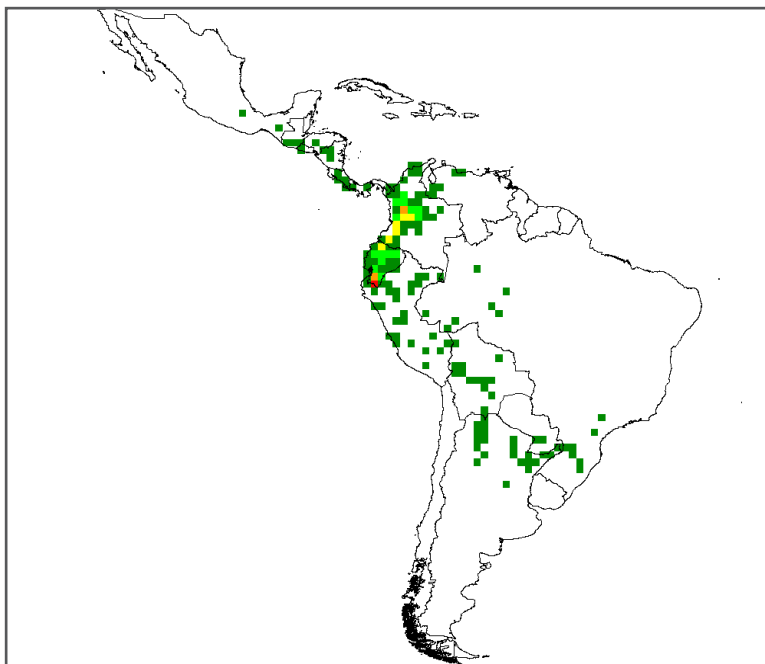
24. Under *Define Grid* window, select the *Use parameters from another grid* option to ensure the parameters of the raster with the observed diversity (the one now being created) are equivalent to those of the raster with potential diversity. To do this, under *Options*, select the raster of potential richness created during the previous steps.
25. Select *Richness* and *Number of different classes (Richness)* as the *Output Variable*.
26. Select the *Simple* option in the *Point to Grid Procedure* box.
27. Select the button to the left of the *Output* box (...) to indicate the name and location of the resulting raster.



28. In the *Parameters* menu, go to the *Field* window and select the item for which you wish to run a diversity analysis. In this case, select *SPECIES*.



29. Click on *Apply* to initiate the calculating process.



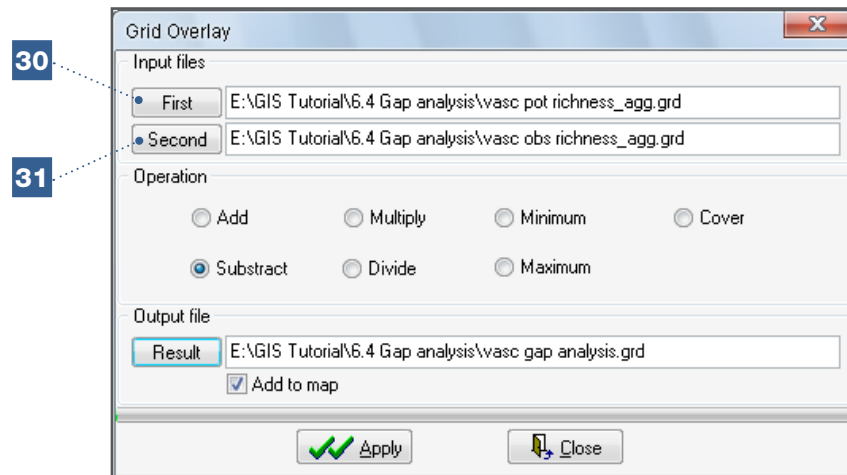
After completing this process, the observed diversity map of *Vasconcellea* species should look similar to the unedited version of the diversity map created in Section 5.1.

Gap identification: comparing potential and observed distribution

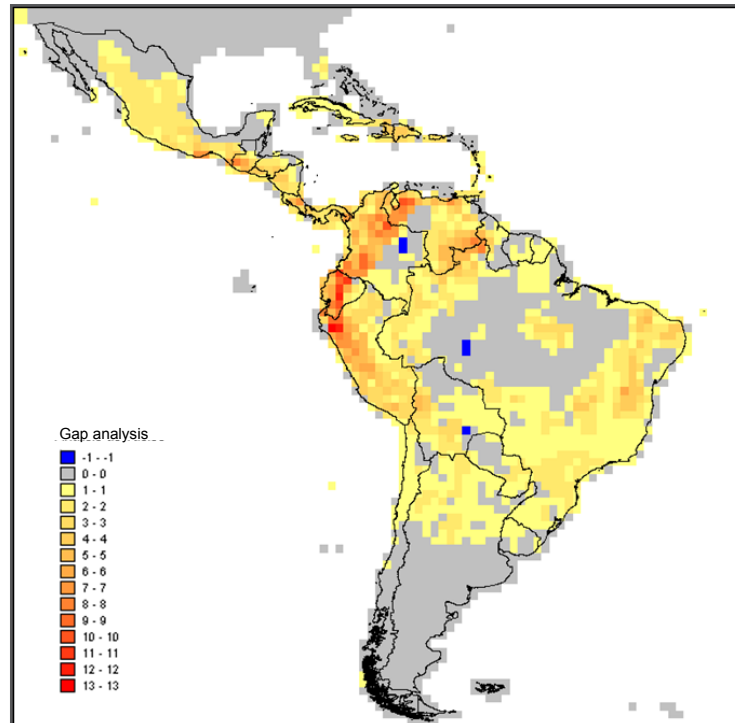
After generating the potential and observed diversity rasters with the indicated parameters, go to the *Grid Overlay* option and subtract the values of the two rasters (similar to Steps 10 to 17 in Analysis 6.3.1).

30. In the *First* window, select the raster of potential diversity.

31. In the *Second* window, select the raster of observed diversity.



Final result:



The map of observed diversity of the *Vasconcellea* species should look like the one presented above. The intense red colour indicates areas where collecting has not yet been conducted, even though these localities are suitable for many *Vasconcellea* species. In fact, in addition to the one or two species observed in these cells, 12 or 13 other species

are expected to be found. These locations occur in northern Peru and in the equatorial transition zone between the Andean and the Amazon regions. Other potential distribution areas for collection are located in Colombia and Venezuela. Certain cells, indicated in blue, have a value of negative one (-1), meaning that more *Vasconcellea* species were observed in these areas than predicted by the potential diversity model.

Note

This section outlines how to conduct a gap analysis for multiple species, but a gap analysis can also be performed for a single species. In that case, you need to overlay the layer of observed and potential richness of a single species (taking into account that the final map will only show the collection gaps and the zones where atypical observations were made).

References

- Anderson RP, Lew D, Peterson AT. 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. *Ecological Modeling* 162: 211–232.
- Araújo MB, Pearson RG, Thuiller W, Erhard M. 2005. Validation of species–climate impact models under climate change. *Global Change Biology* 11: 1504–1513.
- Elith, J, Graham CH, Anderson RP, Dudík M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton J, Mc, Townsend C, Peterson A, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Soberón J, Williams S, Wisz MS, Zimmermann NE. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129–151.
- FAO. 2007. Ecocrop [on line]. Available from: <http://www.ecocrop.fao.org>. Date accessed: October 2010.
- Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27: 861–874.
- Fjeldså J. 2002. *Polylepis* forests Vestiges of a vanishing ecosystem in the Andes. *Ecotropica* 8: 111 – 123.
- Guarino L, Jarvis A, Hijmans RJ, Maxted N. 2002. Geographic Information Systems (GIS) and the Conservation and Use of Plant Genetic Resources. In: Engels JMM, Ramanatha Rao V, Brown AHD, Jacson MT, editors. *Managing plant genetic diversity*. International Plant Genetic Resources Institute (IPGRI) Rome, Italy. pp. 387–404.
- Hernandez PA, Graham CH, Master LL, Albert DL. 2006. The effect of sample size and species characteristics on performance of different species distribution modelling methods. *Ecography* 29: 773–785.
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* 25: 1965–1978.
- Hutchinson GE. 1957. Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology* 22: 415–427.

- IPCC. 2007. Climate Change 2007: Synthesis Report. Cambridge University Press, New York, USA.
- Jarvis A, Williams K, Williams BD, Guarino L, Caballero PJ, Mottram G. 2005. Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. *Genetic Resources and Crop Evolution* 52: 671–682.
- Jones PG, Díaz W, Cock JH. 2005. Homologue: A computer system for identifying similar environments throughout the tropical World. Version Beta a.0. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.
- Liu C, Berry, PM, Dawson TP, Pearson RG. 2005. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography* 28: 385–393.
- Pearson RG, Dawson TP. 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography* 12: 361–371.
- Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modeling* 190: 231–259.
- Phillips S. 2009. A Brief Tutorial on Maxent [on line]. Available from: <http://www.cs.princeton.edu/~schapire/maxent/tutorial/tutorial.doc>. Date accessed: October 2010.
- Puliam HR. 2000. On the relationship between niche and distribution. *Ecology Letters* 3: 349–361.
- Scheldeman X, Willems L, Coppens D'eeckenbrugge G, Romeijn-Peeters E, Restrepo MT, Romero Motoche J, Jimenez D, Lobo M, Medina CI, Reyes C, Rodriguez D, Ocampo JA, Van Damme P, Goetghebeur P. 2007. Distribution, diversity and environmental adaptation of highland papaya (*Vasconcellea* spp.) in tropical and subtropical America. *Biodiversity and Conservation* 16(6): 1867–1884.
- USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network - (GRIN) [Online Database]. National Germplasm Resources Laboratory, Beltsville, Maryland. Available from: <http://ars-grin.gov/cgi-bin/npgs/html/taxon.pl?28462>. accessed: October 2010.
- van Zonneveld M, Koskela J, Vinceti B, Jarvis A. 2009a. Impact of climate change on the distribution of tropical pines in Southeast Asia. *Unasylva* 60 (231/232): 24–28.
- van Zonneveld M, Jarvis A, Dvorak W, Lema G, Leibing C. 2009b. Climate change impact predictions on *Pinus patula* and *Pinus tecunumanii* populations in Mexico and Central America. *Forest Ecology and Management* 257(7): 1566–1576.
- Willis F, Moat J, Paton A. 2003. Defining a role for herbarium data in Red List assessments: a case study of *Plectranthus* from eastern and southern tropical Africa. *Biodiversity and Conservation* 12: 1537–1552.