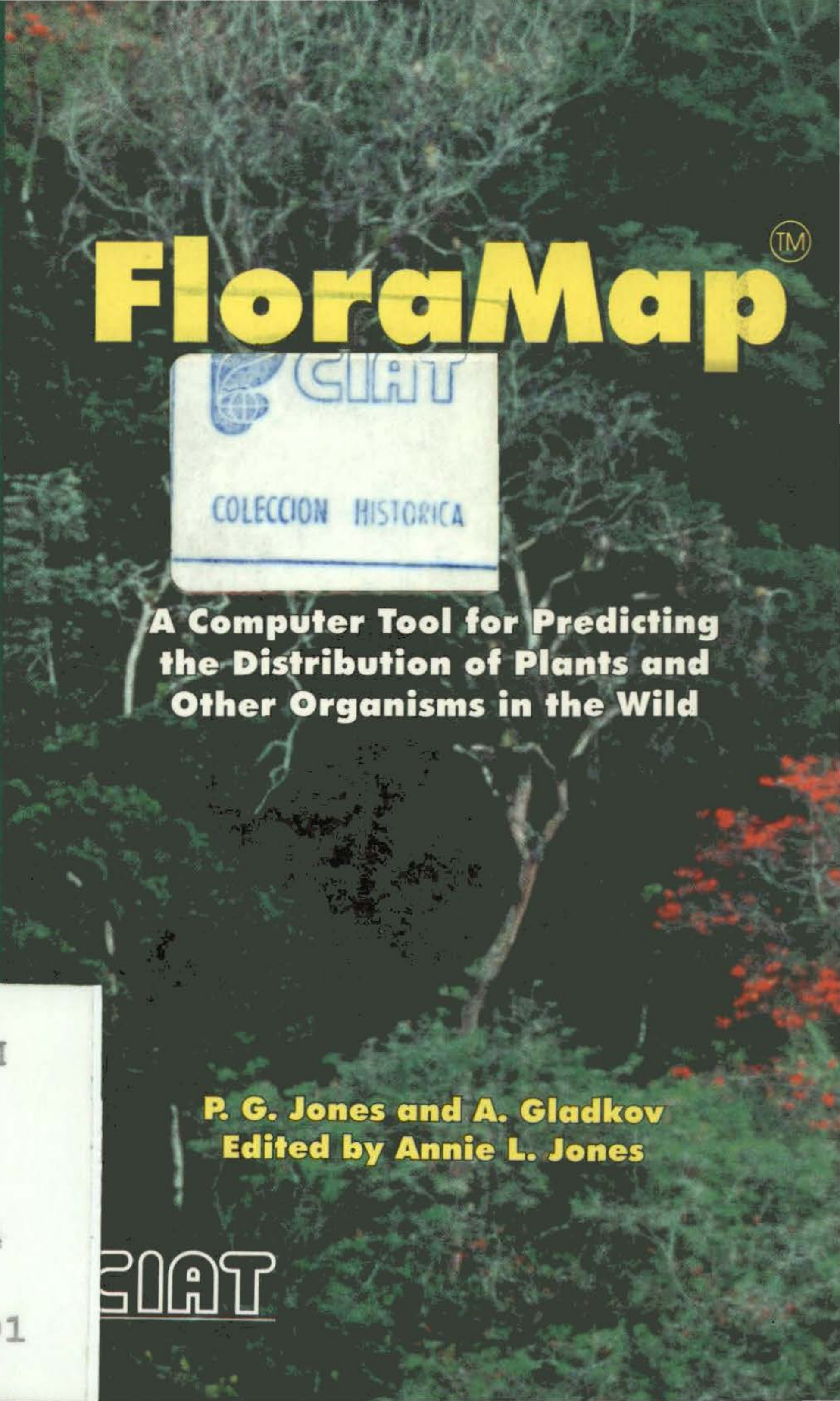


FloraMap™

A Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild

P. G. Jones and A. Gladkov
Edited by Annie L. Jones

CIAT
International Center for Tropical Agriculture

The mission of the International Center for Tropical Agriculture (CIAT, its Spanish acronym) is to reduce hunger and poverty in the tropics through collaborative research that improves agricultural productivity and natural resource management.

CIAT is one of 16 international agricultural research centers, known as the Future Harvest centers, supported by the Consultative Group on International Agricultural Research (CGIAR).

The Center's work is financed by a great number of countries, international and regional development organizations, and private foundations. In 2001, donor countries include Australia, Belgium, Brazil, Canada, Colombia, Denmark, France, Germany, Italy, Japan, Mexico, the Netherlands, New Zealand, Norway, Peru, South Africa, Spain, Sweden, Switzerland, Thailand, the United Kingdom, and the United States of America. Donor organizations include the Asian Development Bank (ADB), the European Union (EU), the Ford Foundation, the Inter-American Development Bank (IDB), the International Development Research Centre (IDRC), the International Fund for Agricultural Development (IFAD), the Kellogg Foundation, the Nippon Foundation, the Rockefeller Foundation, the Wallace Foundation, and the World Bank.

Information and conclusions reported in this document do not necessarily reflect the position of any donor agency.

FloraMap™

A Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild

Version 1.01 2001

P. G. Jones and A. Gladkov
Edited by Annie L. Jones

CD-ROM Series

CIAT

Centro Internacional de Agricultura Tropical
International Center for Tropical Agriculture
Apartado Aéreo 6713
Cali, Colombia

Fax: (57-2) 4450-073
E-mail: p.jones@cgiar.org

Printed in Colombia
November 1999
Reprinted: December 2001 (with corrections and additions)

Jones, P.G. ; Gladkov, A.
FloraMap : a computer tool for predicting the distribution of plants and other organisms in the wild ; version 1.01, 2001. -- Edited by Annie L. Jones. -- CIAT CD-ROM Series. -- Cali, Colombia : Centro Internacional de Agricultura Tropical, 2001.
1 CD-ROM + guide (104 p.)

I. Gladkov, A. II. Jones, Annie L. III. Centro Internacional de Agricultura Tropical.

Contents

Introduction

FloraMap is a system for producing the predicted distribution or the areas of possible adaptation, for natural organisms when little or nothing is known of the detailed physiology of the organism. It is assumed that the climate at the point of collection of a set of individuals is representative of the environmental range of the organism. In the case of plants, these are usually germplasm collection accessions or herbarium specimens.

The climate at these collection points is used as a calibration set to compute a climate probability model. The method uses a Fourier transform to standardize climatic timing and a principal components analysis (PCA) to produce a probability distribution in multiple dimensions. The principal component scores are used to map the probability surface from a set of interpolated climate surfaces. The system has been used to guide plant collecting, to investigate taxonomic and genetic variation, and to map crop pests and their potential predators.

The system is the culmination of over 20 years of work at CIAT and brings together in a user-friendly (or at least non-aggressive) interface some of the techniques that have been developed to cope with CIAT scientists' requirements for these analyses. Peter Jones developed the climate database, the interpolated surfaces, and some of the climate handling functions. Nick Galwey put these together with the PCA in a Genstat package during a study leave at CIAT. This was reported in Jones et al. (1997). Alexander Gladkov put the present system together to run under Windows. We would like to thank CIAT scientists who, over the years, have pushed us to extend the limitations of the system, particularly David Wood, Anthony Bellotti, Steve Beebe, and Joe Tohme. Special thanks are given to Luigi Guarino for his very detailed suggestions, many of which were used to improve the manual.

Using this Manual

The first rule is **READ IT!** Most computer users fall back on the user manual when everything else fails; the authors are no different to any other users. However, the analysis you will be doing in FloraMap is a highly specific type of mapping. It is easy to misapply the algorithms and still produce a pretty map.

Were Microsoft to include a section on the composition of Shakespearean Sonnets in the Word manual, the user would be rightly surprised. FloraMap does not pretend to be a generalized data analysis system and the correct interpretation of the specialized analysis requires a good understanding of the datasets and options used. We hope you will find time to go through the TUTORIAL with us step by step.

You should turn first to the detailed TUTORIAL (Chapter 2) and start working through the example. You can check on the functions of FloraMap in the USER REFERENCE section (Chapter 3) or with the on-line help facility. If you wish to read the theoretical background as you go, by all means turn to the THEORY of the process (Chapter 4). The example is drawn from a real dataset that has been analyzed at CIAT. We have altered some of the accession coordinates to simulate the sort of errors you may encounter in other sets of data.

Once you understand what FloraMap can do, and what it cannot do, you are ready to apply it to your own data. Appendix A will help you to create your input files. **Then you are on your own!** But – do not hesitate to call us if you think that FloraMap is not doing what it is supposed to do.

1. Getting Started

Minimum Hardware and Software Requirements

- **CPU 486 DS, 66 MHz** or better

  Note: FloraMap has been test driven while producing this manual on a Pentium G6 233 MHz. Performance on slower machines is possible but often painfully slow.

- **32 Mb RAM**

- **CD-ROM drive**

- **At least 200-Mb free hard disk drive space**

  Note: This evaluation application will use about ***64 Mb****, the rest is needed for map working space.*

- **15-inch monitor running 256 colors. 1024 x 768 pixels.**

  Note: 16- or 32-bit color is preferred. Under 256 colors, the color range for the probability surface is limited and may appear as texture. If your screen uses less than 768 pixels vertically, you will lose parts of some windows.

- **Windows NT or Windows 95**

  Note: FloraMap has not been tested under Windows 3.1.

- **Color printer or plotter – Postscript preferred**

Installation

FloraMap will install c:\Program Files\CIAT\FloraMap on your C: drive in the directory unless you specify a different path during the installation. You may like to prepare an alternative directory now before you start or you can leave it up to FloraMap installation procedures to create one for you.

✓ Place the FloraMap CD-ROM in your CD drive.

✓ On the run prompt, enter x:\set up where x is the CD-ROM drive device letter.

✓ Answer the prompt questions as the installation runs.

It will create a working directory c:\Program Files\CIAT\FloraMap\demo. This directory contains two accession point files for the TUTORIAL in Chapter 2 of this manual. You may move these to another directory of your choice.

The climate grid files and associated coverage are left on the CD-ROM to avoid cluttering your hard disk with these large files. This means that you will have to leave the CD-ROM in its drive to use FloraMap.

Some users may experience trouble installing FloraMap under Windows 2000. This is because Windows 2000 behaves a little different to NT. Under NT anyone can log on to a machine and, if they are not a registered user, Windows creates a profile for them. Under Windows 2000 it is a little different. A profile is created but it is a virtual one that will disappear when they log off. This affects the privileges the user has. If you cannot load or use FloraMap it probably means you are not a registered user on that machine. Do the following:

- ✓ Log on as administrator
- ✓ Go to Start - Settings - Control Panel
- ✓ click on Users and Passwords (enter)
- ✓ click Add
- ✓ (Browse username on network)
- ✓ click next - [level of access MUST be Standard User (Power User Group)]
- ✓ Click Finish
- ✓ Then log off administrator
- ✓ Log on as user
- ✓ Install Floramap

Join the FloraMap users group listserver to obtain update information and user hints.
Send a message to: **listserv@cgiar.org** with the text:
subscribe FloraMap "your email address."

IMPORTANT NOTICE

If you already have the Borland Database Engine installed you may get the following error message. If BDE is not present, FloraMap will set the value correctly during setup.

"Error accessing climate grids. Too many open files, you may need to increase MAXFILEHANDLE limit in IDAPI configuration."

In this case, open the Control Panel, open BDE Administrator, and follow – SYSTEM and INIT to MAXFILEHANDLES. Increase the number to about 200.

2. Tutorial

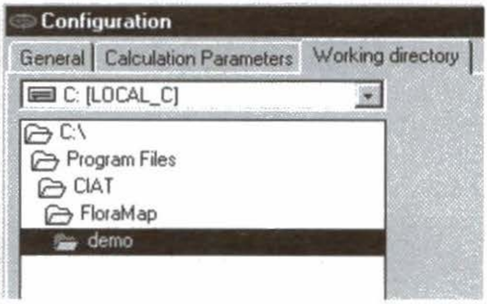
This tutorial aims to familiarize the user with the sort of problems that will occur when using FloraMap on real world data. You can find details of the manipulation of the windows in Chapter 3 of this manual, but we will also include reminders in this section. (Nobody reads manuals until they absolutely have to, do they?) Likewise the theoretical details are in Chapter 4, Theory.

We will work through the tutorial on a set of data that have been prepared from an actual set of accessions of *Stylosanthes guianensis* germplasm from the CIAT germplasm collection. *S. guianensis* is a potentially useful legume for use in tropical pastures. It is widely distributed in Latin America and frequently colonizes areas disturbed by human intervention. The species has a number of different, taxonomically distinct, forms. Attached to each accession, the dataset has an identifier that denotes the cluster assigned from an analysis made in CIAT of an isoenzyme, αβ acid phosphatase. The identifier indicates whether the accession came from group 1 or group 12 in that analysis. Those in group 1 were predominantly *S. guianensis* var. *vulgaris*, while group 12 contained many *S. guianensis* var. *pauciflora* (for further details see Jones et al. 1996).

The accessions data have been altered in certain cases to allow us to show some of the features of the FloraMap system, and to give some practice in chasing down errors likely to be found in comparable sets of germplasm data. The passport accession numbers have all been changed to eliminate confusion with the original dataset. In some cases, a set of imaginary field notes and passport data have been added to lend verisimilitude to the exercise.

Setting up the Map

A working directory has been set up for the purpose of this tutorial. Pull down settings and check the configuration window. The working directory looks as at right.



This directory contains two files. The file stylo_guianensis.dbf contains our tutorial accession points files. The other file, stylo_secondfile.dbf, will be used later in the tutorial to save you time. Accession points files can also be ASCII space-delimited files (see APPENDIX A for preparation of accession points files).

Click on the shortcut to FloraMap icon and we can start. You will see a blank map appear with layers menu in the top right-hand corner. First, we will set the configuration that we will use for the tutorial.

Click on settings and configuration.

- ✓ In the configuration window, click on working directory.
- ✓ Check that the working directory is correct for this tutorial. It should usually be c:\Program Files\CIAT\FloraMap\demo\.
- ✓ Click on general.
- ✓ Set the options autosave configuration, autosave map, and add layer symbols to legend.
- ✓ Check that the sea is blue in map background color.
- ✓ Choose built-in climate grids and select Latin America.
- ✓ Click on calculation parameters.
- ✓ Set on the option show average climate for selected points.

This will show the climate diagram for a selected set of points at any time during the analysis.

- ✓ Set correct temperature off.

Correcting for the elevation of the accession points allows the climate to be estimated with greater precision. We are going to start the tutorial as if no elevation data were available.

✓ Set off treat accessions with identical coordinates as a single observation.

This option changes how the accessions dataset is interpreted. The way you wish it to be interpreted depends on your understanding of what multiple accessions mean for the analysis. Multiple accessions coming from the same coordinate pair can arise in a number of ways. Note here that we are discussing multiple points coming from exactly the same coordinates, not merely from the same climate pixel. This means that the coordinates were entered as identical **on purpose**. Either the collector took a number of accessions from what she considered the same place (the same stop on the road or transect) or the same field or farm; or the sample collected has been subdivided in the germplasm bank processing. This often happens when the sample is grown out to determine phenotypic and/or agronomic characteristics. In the CIAT germplasm bank you can often identify the latter because the accession number is followed by a series of letters (a, b, c, etc.) to denote the subdivision of the sample.

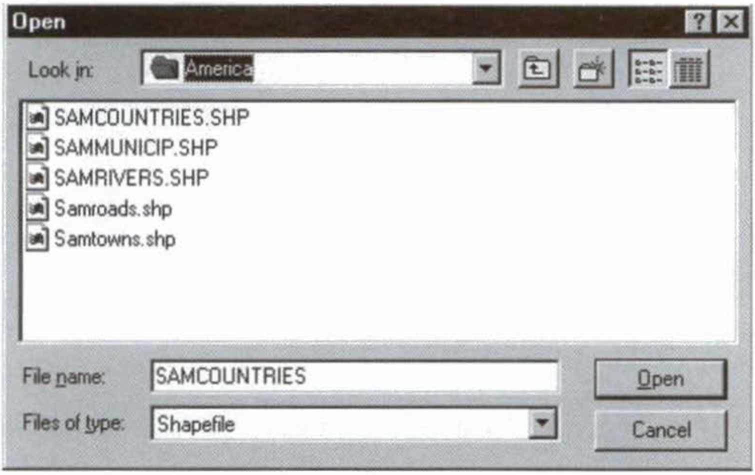
Does this mean that a sample/site is more important because a range of genetic diversity was collected there, or is the sample/site to be considered as one climate point? Either point of view is completely valid and you, the analyst, must decide which to use. In this tutorial analysis we have decided to switch the option off. This means that sample/sites with multiple entries will be weighted heavier than single sample/sites.

✓ Set the mismatches setting to moved manually.

This option is very important. The alternative is to move mismatched points automatically to the nearest point on the climate surface. You will see later in the tutorial that this could produce serious errors if not used correctly.

✓ Click OK to accept the settings and we are ready to go.

The first thing to do is to set up the map. Go to the layers menu and click on the load layer icon. The open layers menu will appear. Choose files of type shapefile. You now have to find the right shapefile to give a background to the analysis. We have placed a selection of shapefiles on the CD-ROM, they are to be found under \COVERAGES\america.



Look in: America

SAMCOUNTRIES.SHP
SAMMUNICIP.SHP
SAMRIVERS.SHP
Samroads.shp
Samtowns.shp

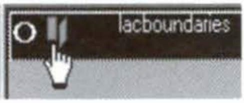
File name: SAMCOUNTRIES

Files of type: Shapefile

SAMCOUNTRIES.shp is the shapefile that will give us the background to Latin America with the country boundaries. There is also SAMMUNICIP.shp, which is the municipal (county) boundaries, and roads, rivers, and towns, which are self-explanatory. All four of these coverages are highly detailed and should not be displayed on the map of the full continent. They are, however, useful when you have the map zoomed in to look at the detail of a region.

☞ *Even when you are zoomed in to a small region, these coverages will be slow to use. The full window of the coverage is calculated even if it does not show on the screen. We hope to alleviate this problem in later versions of FloraMap.*

Highlight SAMCOUNTRIES.shp and click Open. The map of Latin America and the Caribbean will appear in the map window. If it arrives in black, you will need to change the color and fill settings to produce a useful base for your map.



✓ Click on layer control. This will open the layer control menu.

✓ Change the fill color to a good background color. A dull green is quite effective.

✓ Set to solid fill.

✓ Set off show in legend, you do not need the background named.

✓ Set on with outline.

✓ Set outline color to a dark color, black will usually do.

✓ Click on apply to check what you have done. If it looks good, click on OK to set the changes.

Now you are ready to load the accessions dataset.

✓ Click on the load layer icon, as above.

✓ Choose accession points as the type of file.

☞ ***NOTE:*** *Because the file extension .dbf is used also for the shapefiles, these will also appear as accession points. Unfortunately, because of their geographic contents, FloraMap can accept them as valid accession point files. Beware of opening these as accession point files because they are often very large and will take an inordinate amount of time to load. We are working to make this aspect safer.*

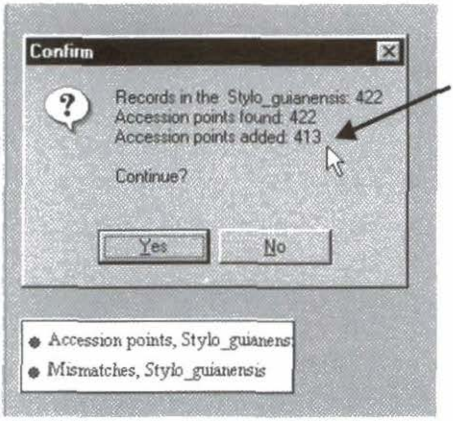
✓ Change the directory back to the working directory c:\Program Files\CIAT\FloraMap\demo.

✓ Select the file stylo_guianensis.dbf and click open.

FloraMap will open the file and proceed to check the contents. FloraMap does this by ensuring that every point in the file can be allocated to a valid climate pixel from the climate database. This is a long operation as the whole database has to be scanned to determine the validity of the points. The words checking accessions will appear at the bottom of the map window in the process indicator and a timing scale will appear. Because this operation is a top-level activity, the layers window will not disappear from the screen if you try to switch into another application while the accessions checking is in process. To avoid this interfering with your other work, move the layers window to an unobtrusive part of the screen before opening the accession points file.

Checking the Data

The accession point checking has now finished and you can see the following window.



Confirm

Records in the Stylo_guianensis: 422
Accession points found: 422
Accession points added: 413

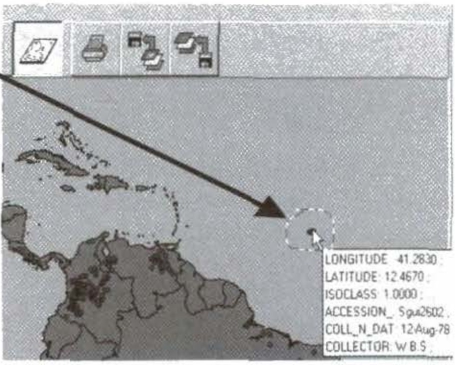
Continue?

- ● Accession points, Stylo_guianens
- ● Mismatches, Stylo_guianensis

The checking did not find all of the accession points. Luckily, you selected the manual movement of missing accession points. Now you have to check the map and find out where they are. There are 422 - 413 = 9 points to be accounted for.

✓ Check no on the confirm window. Release the confirm and layers windows to release space to manipulate the map, and go in search of the missing points. The map legend indicates they will be shown in red.

Here you have found a group of accessions all alone on the Atlantic Ocean.



LONGITUDE: -41.2830 ;
LATITUDE: 12.4670 ;
ISOCLASS: 1.0000 ;
ACCESSION_: Sgui2602 ;
COLL_N_DAT: 12-Aug-78
COLLECTOR: W.B.S ;

✓ Use the zoom and pan until you can see the accessions clearly.

✓ Draw round the points with the area select tool. *Make sure that the mismatches dataset is highlighted in the layers control window.*

✓ Or point at the accessions with the cursor until the information appears.

This is a classic case of mislocation. It is almost always correctable by the slightest of clues. In this case, look at the collector's identity and when the accessions were collected.

W.B.S. was the collector and she collected these accessions on 12 Aug, 1978. You can be quite confident that she was not in the Atlantic on that date, so where was she? You have some options.

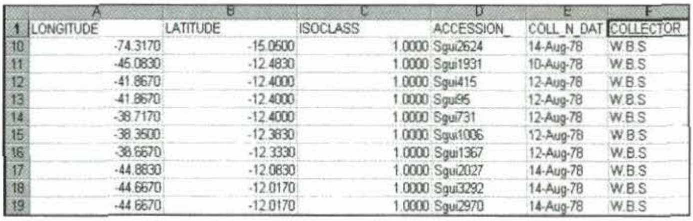
✓ Browse through the accessions set in dBase, Excel, or other spreadsheet looking for the date and collector.

✓ Check back to the germplasm database that the accession set was taken from to see if there are other clues.

✓ Check to see if there is a published catalogue of the collection.

✓ Check with the original fieldbook of the collector if you have it.

In this case, we have access to two forms of the information. The accession set list shows us a group of accessions in Brazil collected by W.B.S. around that date.



	A	B	C	D	E	F
1	LONGITUDE	LATITUDE	ISOCLASS	ACCESSION_	COLL_N_DAT	COLLECTOR
10	-74.3170	-15.0600	1.0000	Sgui2624	14-Aug-78	W.B.S
11	-45.0830	-12.4830	1.0000	Sgui1931	10-Aug-78	W.B.S
12	-41.8670	-12.4000	1.0000	Sgui415	12-Aug-78	W.B.S
13	41.8670	-12.4000	1.0000	Sgui95	12-Aug-78	W.B.S
14	-38.7170	-12.4000	1.0000	Sgui731	12-Aug-78	W.B.S
15	-38.3500	-12.3830	1.0000	Sgui1006	12-Aug-78	W.B.S
16	38.6670	-12.3330	1.0000	Sgui1367	12-Aug-78	W.B.S
17	-44.8830	-12.0830	1.0000	Sgui2027	14-Aug-78	W.B.S
18	-44.6670	-12.0170	1.0000	Sgui3292	14-Aug-78	W.B.S
19	-44.6670	-12.0170	1.0000	Sgui2970	14-Aug-78	W.B.S

Now you know that W.B.S. was in Brazil at 12.4 South, 41.867 West on 12 Aug, 78. It is therefore likely that the accessions came from there. Your second piece of confirming information comes from the notes field of the passport data in the germplasm database. They state:

12th Aug between Itaberaba and Seabra, Paraguacu valley, foothills of Serra do Sincora. The evidence therefore points to the case of the missing sign. In this instance, on the latitude. Because the corrected coordinates put the mismatched accessions directly on the road from Itaberaba to Seabra, we can assume that this was the only (unfortunately common) error in the data.

✓ In your database or spreadsheet program correct the latitudes for these points in the accession points dataset.

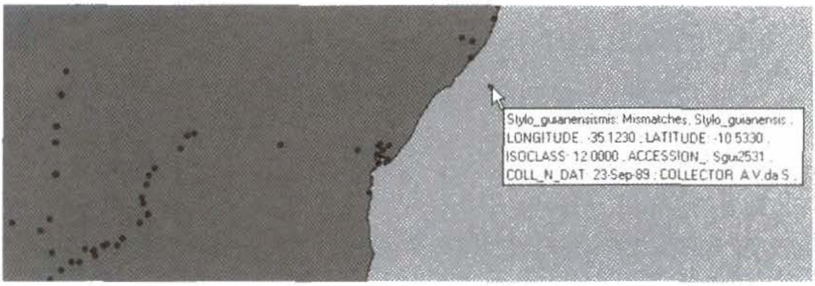
☞ ***NOTE:*** *Remember to close the accessions table after you have changed it. FloraMap will not be able to access it if you leave your spreadsheet in a minimized window.*

There is another single point out to sea, off the coast of Brazil.

✓ Choose the zoom tool and close in on the east coast of Brazil.

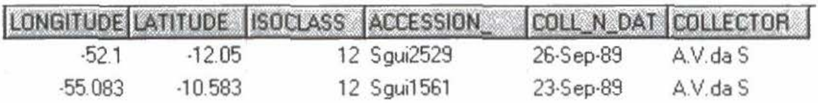
✓ Using the panning icon, move the map up and to the left to bring the area into easy working range.

✓ Point at the red mismatched point until the information appears. Or draw around it with the area select tool.



Stylo_guanensismis: Mismatches, Stylo_guianensis ,
LONGITUDE -35.1230 , LATITUDE -10.5330 ,
ISOCLASS 12.0000 , ACCESSION_ Sgui2531 ,
COLL_N_DAT 23-Sep-89 ; COLLECTOR A.V.da S .

This time you have an accession collected by the famous (and fictional) Brazilian botanist A. V. da Silva, as you can see from the information on the accession point. You need to find out where he was on 23 Sept, 1989. You check in the accessions list and luckily find that in Sept 1989 he was collecting in Mato Grosso do Norte where the *Cerrados* meets the Amazon forest.



LONGITUDE	LATITUDE	ISOCLASS	ACCESSION_	COLL_N_DAT	COLLECTOR
-52.1	-12.05	12	Sgui2529	26-Sep-89	A.V.da S
-55.083	-10.583	12	Sgui1561	23-Sep-89	A.V.da S

The germplasm database does not carry any more information on these accessions. Fortunately, Dr da Silva published the journal of his collecting trips from 1956 until he retired in 1991. From these you can read the following descriptions:

> *Setem 26 Varzias do rio Xingu, Floresta de varzia muito denso, poucos leguminos baixos. Uma S. gui.*
> *Setem 23 Noreste de Telez Pirez. Floresta de cerradao denso. Mata de varzia em baixos, dois S.gui em clareiras dos acampamientos. Caminho do varios kilometros.*

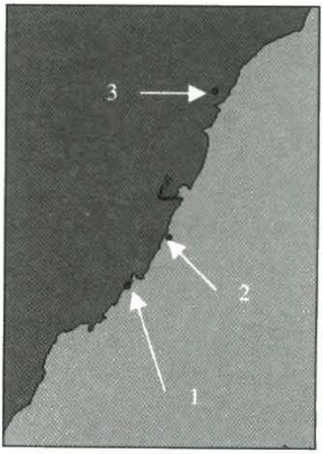
This puts the accession in the correct area, northeast from Telez Pirez. But what are the correct coordinates? Neither latitude nor longitude match the other accession site. However, we note that the collector was moving locally between campsites (miners' perhaps?) so the other site should not be too far away. Could a simple data entry error change these coordinates into a place a few kilometers from the other? Yes, in this case a common keying error has replaced a 5 with a 3. In fact, location 55.083 west, 10.583 south is exactly 5 km 460 meters from 55.123 west, 10.533 south. 55.123 west had become 35.123 because of the keying error.

✓ If you agree that this evidence is convincing, go ahead and correct the point in the accessions list.

The next problem is one that FloraMap is designed to deal with. When the climate database surfaces are computed, we rely on a digital elevation model (DEM) of the area. This is a computer generated representation of the topography, in this case a simple square grid giving the modal heights of the land in each grid cell. This, in turn, relies on maps at various scales and projections.

The background and key geographic layers incorporated in FloraMap also come from maps on various projections. The maps used by botanists and plant collectors have the same characteristics.

This makes it relatively likely that, at the continental scales at which we are working, a slight mismatch sometimes occurs at the edges of the coverages. This usually results in a few points falling off the edge of the climate coverage.



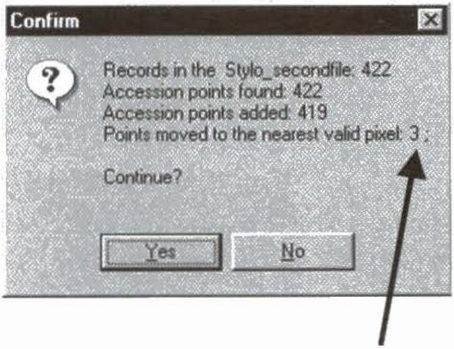
Look now at the coastline north of Rio de Janeiro. Point 1 appears to be so close to the coast it might almost be on the beach. (Questions as to why plant collectors are idling on Brazilian beaches should be addressed to the relevant authorities and not to the authors of FloraMap!) Point 2 also appears to be close to shore and a small nudge would put it onto a valid pixel in the climate database. Point 3 appears to be a conundrum. It is on land, but remains a mismatched accession. Here you have to remember that the climate database pixels are about 18 km on a side. Point 3 is in the corner between two pixels that are approximating the coastline at this point. It is usually wise to check the coordinates of points that you find in the same situation as these three, in case you can spot any obvious errors. However, all you have to do after that is to restart the analysis after changing the configuration.

- ✓ Clear the map.
- ✓ Click on settings.
- ✓ Choose calculation parameters.
- ✓ Under mismatches choose moved automatically.

☞ ***NOTE:*** *Do not do this until you have finished checking all mismatched points. If you leave points unchecked, FloraMap will very kindly put them at the nearest point on the coastline whether or not it is the sensible thing to do!*

In this case you have, so click OK and proceed with the tutorial.

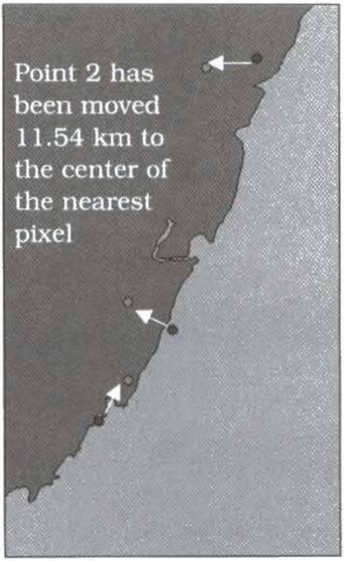
Restart now.



Records in the Stylo_secondfile: 422
Accession points found: 422
Accession points added: 419
Points moved to the nearest valid pixel: 3 ;

Continue?

Three pixels are no longer mismatched. They have been moved to the nearest valid pixel.


Point 2 has been moved 11.54 km to the center of the nearest pixel

These points have been added to the end of a new file of 422 accession points that FloraMap will use as of now. You can delete the mismatched points set from the map if you wish. They will take no further part in the analysis.

It may have occurred to you that if you can see obvious errors, such as mismatched points, there may be errors that have not translated points off the working layer of climate, but merely moved them to another valid, but wrong, climate pixel. This, unfortunately, is usually the case.

An obvious check is to verify that the points are at least in the countries and/or regions where you know the accessions were collected. In the case of these *S. guianensis* collections we could safely say that any points falling in Argentina or Chile would be out of bounds. But this is relying on the knowledge that none of the collectors went to those countries, not necessarily that it is impossible to find the plants there.

Another check, if accessions are well-dated, is to actually follow out the collecting trip on the map. The FloraMap roads' coverage, which you will find on the CD-ROM, can assist with this. Unusual jumps or sidetracks in the itinerary need to be investigated. The distance tool (p 51) can quickly indicate if a sidetrack could feasibly be made in the time.

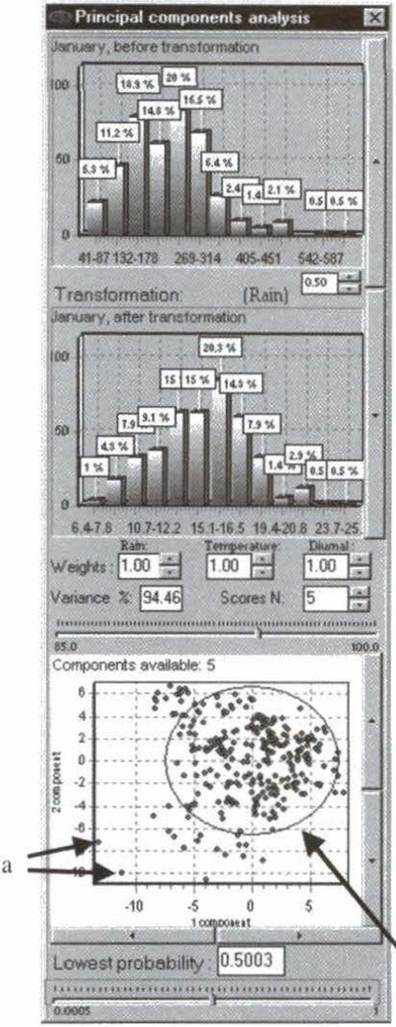
When all of these obvious methods have been exhausted, FloraMap still has a few tools that may help. Carry on with the analysis and you will get a chance to apply them.

✓ Click on yes on the confirm menu at left.

The message creating climate file will appear in the process indicator, along with its timescale bar. Once this is finished, the PCA window will appear. This is a very full window, please take time to have a look at its description in Chapter 3.

We are still primarily interested in checking the dataset. In a real analysis, this can take days or even weeks of your time. People tend to think that data become better for being written down and that once inside a computer they are 'truth'. Germplasm passport data may have passed through a dozen or more stages in their processing. Coordinates go from map to notebook, from notebook to report. Often, the report goes through a publisher's hands. This involves editing and typesetting. The data are then coded for database entry, and, as accessions are passed from collection to collection, the coded data are transferred from one computer system to another. It is a wonder we get anything at all that is worth working with, but because all the people involved are highly dedicated, we usually do. Mistakes, however, are inevitable.

Let us skip straight to the next level in the data checking.



Principal components analysis

January, before transformation

Transformation: (Rain) 0.50

January, after transformation

Weights: Rain: 1.00 Temperature: 1.00 Diurnal: 1.00

Variance %: 94.46 Scores N: 5

Components available: 5

Lowest probability: 0.5003

Look at the PCA window as at left.

- ✓ Set weights to 1.00.
- ✓ Set transformation to square root - rain to the power 0.5.
- ✓ Set scores N to 5.

You should have the analysis looking like the picture. If not, make sure that the corrections you made to the accessions dataset were accepted correctly.

Look at the PCA scattergram at the bottom of the PCA window. This one is showing the plot of the accessions in the two-dimensional space defined by the first and second principal component. If yours is not, toggle the side bars until you have the correct diagram displayed. This component-1-by-component-2 plane is almost always the plane in the PCA space that shows most of the variation in the data. In fact, in this analysis they account for about 70% of the variance.

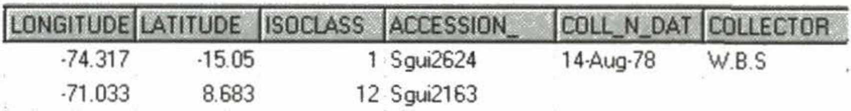
The ellipse is showing the boundary at 2 standard deviations. In two dimensions we expect about 14% of the points falling outside this ellipse. With 422 accession points we would therefore have 49, so the population is not seriously non-normal. However, a few points deserve investigation.

The two points at the bottom left corner (a) are obvious outliers. Check where they are on the map by using the area select tool to draw around them. The accessions should look like the two below, Sgui2624 and Sgui2163.

Now wait a minute! We have already seen that W.B.S. was collecting in Brazil on 14 Aug, 1978, so how can she also be in Peru? Unfortunately, in this case we do not have access to her fieldbooks and the passport data contain no helpful notes. We know roughly where the other accessions were that she collected on that day, but the missrecorded coordinates hold no helpful hint. There is no obvious slip of the keyboard, someone must have skipped a line or shuffled sheets when these data were recorded. This case has beaten us. At least we know it is wrong, but the only recourse is to delete the accession.

✓ Exit from FloraMap, go to your spreadsheet and delete the row for accession Sgui2624.

And then there were 421.



LONGITUDE	LATITUDE	ISOCLASS	ACCESSION_	COLL_N_DAT	COLLECTOR
-74.317	-15.05	1	Sgui2624	14-Aug-78	W.B.S
-71.033	8.683	12	Sgui2163		

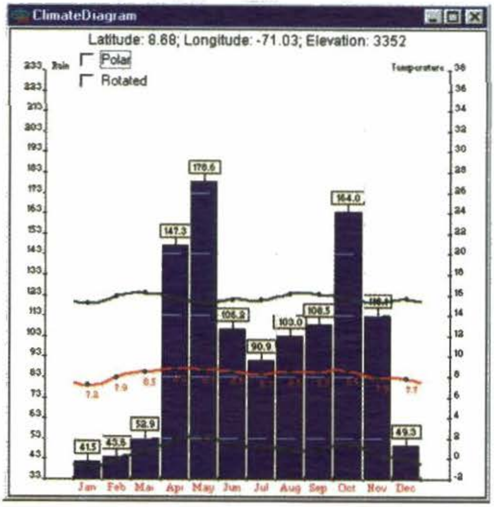
Look at the next point, Sgui2163. It is in Venezuela. There are no notes, no fieldbook, and no friendly date and collector. (This is not actually true because in real life it is a CIAT-collected accession and we would know its exact provenance if it had its correct accession number. In fact, this is the first case of a problem that has not been invented that we have come upon in this exercise.)

So what is making the climate of this accession so different? Take a look at the climate record that was generated for this accession from the database.

✓ Draw around the point on the scattergram with the area select tool. You have got the calculate average option set on, so a climate diagram will appear. Ignore it for the moment. Notice that the point in Venezuela starts flashing on the map. You have identified where it is!

✓ Zoom well in to enlarge the area round the point. You can take it until no other points are showing.

✓ Use the climate diagram tool to point at the accession. The climate diagram will appear as at right.


ClimateDiagram

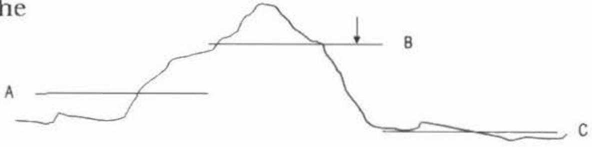
Latitude: 8.68; Longitude: -71.03; Elevation: 3352

□ Polar

□ Rotated

It would be a very strange *S. guianensis* plant that would grow in average daily temperatures of around 8 degrees centigrade, with night temperatures falling below zero in at least two months. What has happened? Look at the elevation of the climate grid pixel. It is 3352 meters, well up, high in the cordillera.

This problem arises because of the 18-km size of pixel used by the climate database. However, it will not be completely resolved with a closer grid of climate. Remember the problems of mismatch along the coastline? Exactly the same thing can happen, for all the same reasons, along a topographic discontinuity. This could be a line of hills, an Andean cordillera, or the edge of the rift valley in Kenya. The climate is generally insensitive to a few hundred meters, or even a few kilometers, of lateral displacement. Slight mismatches of map layers do not usually matter much for rainfall. But, if the mismatch changes the elevation, the temperature can be far different to that expected.

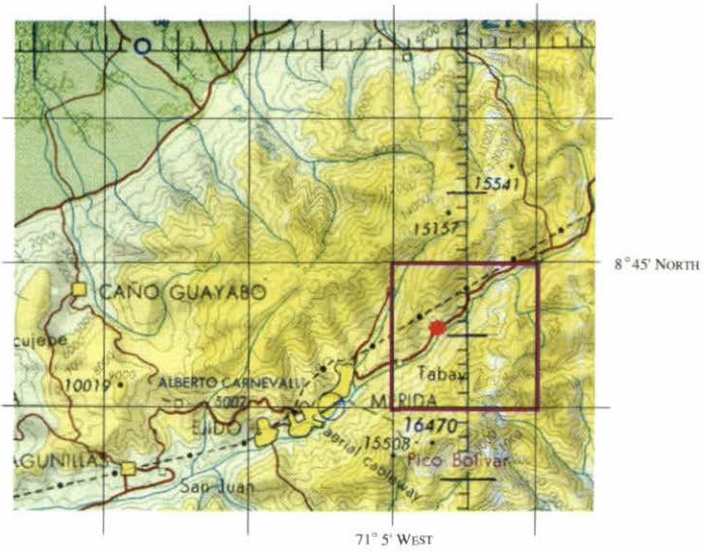


Each of these pixels, A, B, and C, represents the modal elevation of the terrain it covers. In the database we are using they are 18 km long. The presence of a significant hill cutting up pixels A and B is no surprise at this scale. If an accession point is represented by the arrow on pixel B, what is the solution? We could move it to pixel C as we would have done if it had missed a coastline. That is a viable solution if you know the terrain and know that moving to pixel C would be valid. A small shift of coordinates would do the trick.

However, suppose the collector really did collect the specimen from the point on pixel B and no displacement has occurred. The elevation at that point is well represented by that on pixel C, but we do not want to shift the point. We can correct the climate record to the elevation at the collection point. The lapse rate correction for temperature is simple and reliable. Temperatures drop roughly 6 degrees for every 1000 meters of elevation (FloraMap uses a more precise model, but the idea is the same).

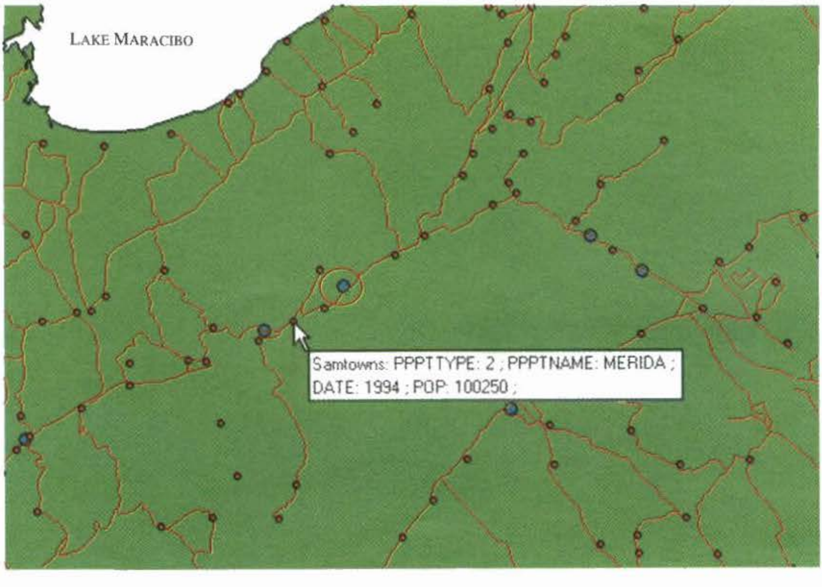
If FloraMap is to do this, it needs to know the actual elevation at the collection point. We have written FloraMap to be flexible on this point because (especially from old collections) we do not always have the origin elevation as measured by the collector. If no elevation data are available, you will start the analysis with an accessions dataset with no column for elevation. FloraMap will accept this and fill in with elevations from the climate database. In fact, as the temperatures in the database are adjusted to the pixel elevation, there is nothing to do. However, if you put in a column labeled elevation, FloraMap will look down this column and anywhere a valid nonblank field occurs, it will substitute that elevation for the one in the database and automatically correct the temperature with the lapse rate model.

Here is a section of the 1:1 000 000 Operational Navigation Chart (ONC) K 26 for the region around Mérida in Venezuela.



☞ *These charts are available for the whole world. They are published by the Defense Mapping Agency Aerospace Center, St Louis AFS Missouri 63118. They are reasonably reliable and cheap. Air navigators use them to draw courses on, so they have to be cheap enough to be disposable. They are an indispensable aid to the geographic detective work that you will need to do to validate your accessions datasets.*

The climate pixel grid is shown as the black graticule. The chosen pixel for accession Sgui2163 is shown in purple. The pixel coordinate denotes its northwest corner and the pixel spans 10 minutes as the map graticule shows. The red star shows that the plant was collected (as are many accessions of certain species) by the side of the road. This is a useful check on the validity of the coordinates. You can see why the pixel elevation is over 3300 meters. The dark yellow area is all above 9000 ft (unfortunately the ONC charts are not metric) and this dominates the landscape in the pixel with relatively large areas exceeding 12 000 ft. The pixel is therefore an accurate generalization of the landscape, but Sgui2163 came from the bottom of the valley, close to the river, at a little less than 7000 ft. This translates to about 2100 meters, just within the limit for the species.



Samtowns: PPPTTYPE: 2 ; PPPTNAME: MERIDA ; DATE: 1994 ; POP: 100250 ;

Here is an alternative way of viewing the area from within FloraMap. If you want to try it, follow these steps:

✓ Zoom in to the area shown.

✓ Select samroads from the coverages directory on the CD-ROM.

✓ Open the layer control icon for samroads in the layer control window.

✓ Set size to 1 and the color to red.

✓ Select samtowns from the coverages directory.

✓ Set size to 3 and color to orange.

✓ Change the accession points size to 6 and color to blue.

Now we have to make the correction for elevation.

✓ Go to your spreadsheet.

✓ Open the S-guianensis.dbf file and insert a column for elevation.

✓ You can now enter 2100 against Sgui2163.

At this point, you could rerun the analysis and note that the point for this accession would disappear from the outliers in the scattergram and join the others inside the ellipse. However, some accessions come from similar Andean situations. It will be worth checking these others for the same type of problem.

✓ Use the climate diagram tool to point at a likely candidate. We know the Andes are high, near the Colombian-Ecuadorean border. Try the points at 77.5 degrees west, 0.9 degrees north. It turns out to be accession Sgui1285.

The climate diagram gives the elevation at 2743 – clearly not an elevation at which we should find *S. guianensis*. It would be best if we can give FloraMap the elevation of all the accession points so that we do not have to chase all the possible errors one by one.

To save you time and effort (although there is no shortcut when you come to real data!) we have prepared a new accessions list for you, incorporating the elevation data for all the Andean accessions. Restart the analysis with stylo_second file and you will have the collectors' estimates of elevation, and all the corrections that you have incorporated so far.

The PCA Analysis

✓ Either erase all accession points layers from the map, or restart FloraMap.

✓ Pull down configuration from the settings menu.

✓ Select calculation parameters.

✓ Select correct temperatures.

✓ Under mismatches select moved automatically.

☞ *It is important that you use this last option only when you are sure that the remaining points to move are those that are grazing the climate dataset boundary. If you use this option when points are falling far outside the boundary, they will be moved automatically to the nearest boundary. This is usually an error.*

✓ Click on OK to reset the configuration.

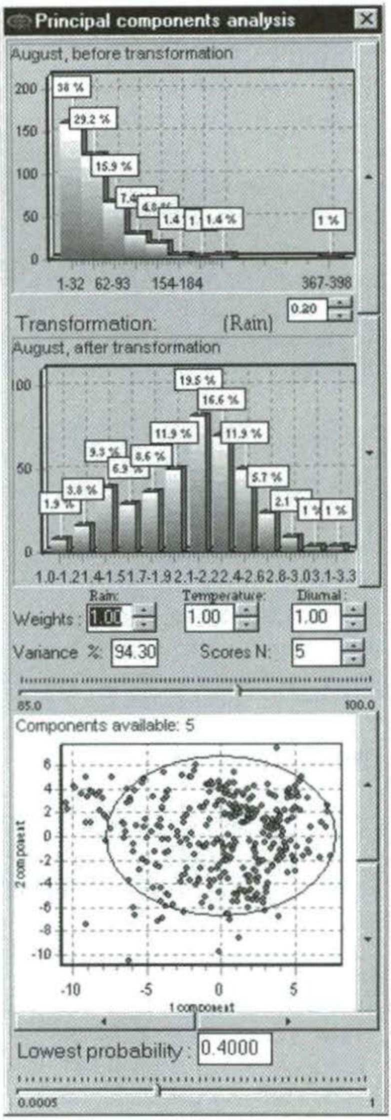
✓ From the layers menu select accession points and then Stylo_secondfile.dbf.

✓ At this stage you can delete the mismatched layer from the layers control window. These points have now been corrected and added to the accession points file. To ensure that they are recorded, save the accession points file from the layers control window.

✓ Click on the PCA icon to start the new analysis.

The PCA window will appear as at right. We are going to use some of its features to trim the following analysis. First, set the transformation for the rainfall data; they are notoriously non-normal, unlike temperature data, which are almost always well behaved. They are often modeled as a gamma distribution (see Jones and Thornton 1993, 1997). Because FloraMap uses the special characteristics of the normal distribution to determine probabilities, you need to select a transformation that will change the data's distribution to as closely approximate a normal distribution as possible.

FloraMap presently offers two transformations well suited to rainfall data—the natural logarithm, ln (rain), and the exponential, rain^x. In the former case, only one option exists. In the second,



Transformation: (Rain) 0.20

Weights: Rain: 1.00 Temperature: 1.00 Diurnal: 1.00

Variance %: 94.30 Scores N: 5

Components available: 5

Lowest probability: 0.4000

exponents range widely, from –3.0 to -0.1 and from +0.1 to +3.0. The middle range is disallowed for computational reasons (small positive or negative exponents transform all values too close to 1.0 to be of any practical use). Rain$^{0.5}$ is, of course, the square root transform. Rain^{-1} is the reciprocal. Changing the exponent in the window provided can specify many other useful transforms. Common transformations such as the logit, probit, and trigonometric transforms are not included in FloraMap because they are of dubious use with rainfall data.

✓ Click on transformation to toggle between the logarithm and the exponential transformation.

Watch how the choice of transformation changes the distribution in the scattergram window. As you change the exponent, the number of scores selected will occasionally change. This is because the algorithm is taking the percentage of variance explained as a constant and tries to adjust the number of scores to compensate for the changing data. Some of the transformations you can achieve with the exponent transform are unusual to say the least, but, even if never used for a mapping, they can give an interesting picture of the groupings of data in the scattergram.

The transformation should produce, as closely as possible, a typical bell-shaped normal curve for each month (scroll through the months with the scroll bars at the side of the histograms). At present, we cannot fit individual transformation to each month; a single one has to suffice for the whole year. The requirements of a good transform cannot usually be satisfied simultaneously for all months. Those with least rainfall are particularly hard to normalize. In this case these are July, August, and September. Use the transformation tools interactively to accustom yourself with what they are doing, and choose the one that gives the best results over the whole year. Unfortunately, this may not be the one that fits best to some months. The one illustrated here is rain to the power 0.2 or the fifth root of rainfall. Verify that this gives the best fit to all months, that June is the worst fitted, and that any other power introduces more distortion.

☞ *We hope that future versions of FloraMap will incorporate statistical aids for selecting transforms and further generalized transformations.*

Now that you have selected your transformation, you can proceed to the next step. We will leave the weightings at 1 for now and look at the effect of choosing different numbers of principal components.

The Probability Analysis and Mapping

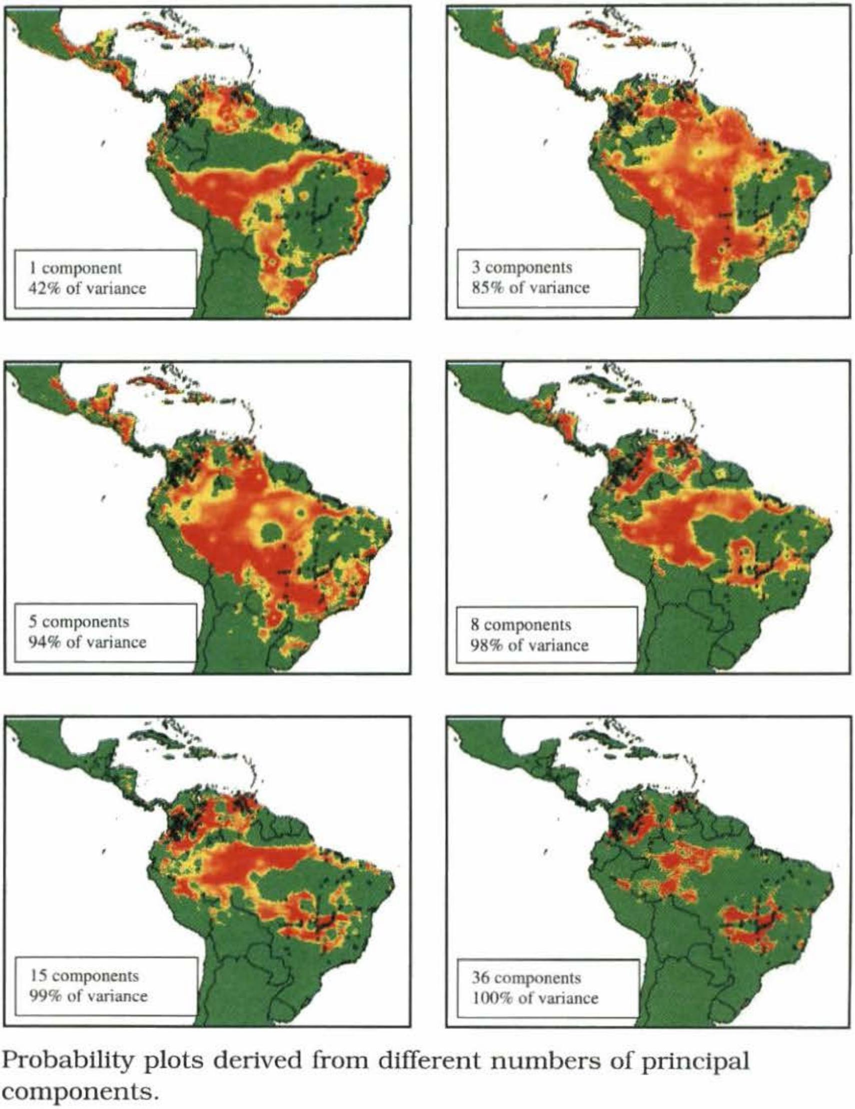
We are now ready to start mapping the probabilities. What we will map is the probability that each pixel on the map, and hence in the climate data grid, belongs to the population of climates described by the PCA analysis that is presently in the PCA window. See the THEORY section, p 68, for details.

✓ Check the scores N box and toggle to the number of components that you would like to use. Try out a few times with different numbers of components to see what happens. You can perform the same operation by moving the variance lever or by stipulating the desired variance in the variance box. Notice that a range of variances apply to each set of components (scores N) selected.

✓ Each time, use the map display icon to produce the probability layer. Set the minimum probability to 0.4 to give a close coverage of high probability points. You can delete old layers from the layers menu if you want to retain an uncluttered map.

✓ You can save the probability layers that look interesting, giving them a new name and recalling them as rendered shapefiles from the layers menu.

The images below show the effect of using different numbers of principal components in the present analysis with weights set to 1 and lowest probability to 0.4.



Probability plots derived from different numbers of principal components.

The maps with few components fit to large areas outside the range of the accession points and show surprisingly poor fit in the main accession areas. As the number of components is increased, the probabilities fall more into line with the accession points, and areas outside the main collecting areas are diminished in importance. This is not surprising. The first few principal components measure very broad, overall characteristics of the climate. The first is usually a **size**-type of component. Variation along its axis will often be associated with, for example, more/less rain, higher/lower temperatures. Thus, it is a crude measure of the climate, but one that explains much of the variation (42%). As more components are added to the analysis, the proportion of the variance accounted for rapidly rises.

By the time five components are included, they are accommodating 94% of the variance. Only a meager 6% remains for all the others. This would normally be considered noise, left over from the recording and processing of the climate data. We could consider fitting up to eight components, which would take the fit to such a high level that only 2% of the variance is unexplained. This must definitely be noise. Temperatures are only measured to 0.1°C and certainly not so accurately in the long term.

Then why does the fit of the probability area seem to increase in accuracy right up to the fit with 36 components? The answer is two-fold. First, the noise has sufficient random coherence that the fit appears to be refined, well beyond the point when random noise should be taking over. Second, each additional component carries with it 13/12 degrees of freedom from the population parameter estimates.

Thus, if we were to fit a calibration set of 36 observations, the map of the probability surface for 33 components would by definition fit exactly to the calibration points. In fact this is what happens, each accession point is close to a local maximum on the probability surface. But no data are free to produce a valid interpolation surface. We therefore do not want to fit too many principal components, especially with small calibration sets. Exactly how many to choose is not an exact science, the fewer the better, but they have to produce a realistic probability surface. Normally, we would prefer to use somewhere between four and eight, depending on the size of the calibration set.

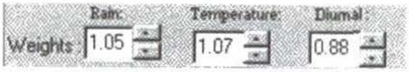
☞ *FloraMap will not let you fit more components than the data will theoretically allow. However, you can still fit more than the information content of the data warrants.*

Unfortunately, when we fit even up to eight principal components in this dataset, large areas of high probability remain where no calibration points fall, and large areas where accession points have very low probability. Having positive probability areas without calibration points may not be a bad situation. They may well be correctly predicting the possible existence of germplasm that has not yet been collected. The converse, where the map is showing low probability for areas with accession points, is obviously wrong. For a possible explanation of what is wrong, see THEORY section, p 78.

If we allow the arguments put forward in the Theory section, then we are dealing with more than one climate population distribution. We do not, at this point, have any evidence apart from the probability fit on the map. We have no idea of how many climate populations may be hidden in the data of the calibration set. Least of all do we have any evidence that the different climate populations reflect genetic variation in the species that we are mapping. FloraMap incorporates a tool for investigating these possibilities from the climate data. We strongly advise you to now return to the germplasm data and see if any information can be gleaned there that may be useful in separating possible populations. You will see a simple example of this later in the tutorial. In practice, it is wise to start thinking about the problem from both sides as soon as possible.

A last option that we should cover in the probability analysis is the weighting of the variates. Until now we have left the weights at 1. The 36 variates are grouped in three groups of 12. You have already seen how the transformation is applied only to one group of the variates (rainfall). The weightings are applied to the full dataset in groups of 12. There are therefore three weights, one each for rainfall, temperature, and diurnal temperature range. They add to 3.

To change the weights look on the PCA window where the weight boxes and toggles appear. Although we have not looked at the cluster tool yet here is a useful hint. It is best to turn off


Rain: Temperature: Diurnal:
Weights 1.05 1.07 0.88

clustering when you move the weights. Each time you change a weight, the clustering is recalculated. This can result in cumbersome waiting time. Of course, if you want to see the effect of the weight change on the clusters then you should recluster after you have changed the weights. In this case, it is preferable to enter the weights box and type in the new weight directly. If you try to toggle the weights using the levers, both the PCA and the clusters are recalculated for each point that you pass.

✓ Toggle the rain weight up as far as it will go.

See how the scattergram responds with clusters concentrating and disappearing. Points are scattered off the surface of the distribution ellipse. You will see similar behavior if you toggle through the weights for temperature or diurnal temperature range. You will probably never use a weighting of 3 for rainfall, but the option is highly useful as you scroll through interactively increasing the weight. The dynamic action of the points can often give you an idea of those that are special. For a data point, being special unfortunately often indicates an error. Strange behavior should be investigated. At the minimum, it will lead you to identify unusual environments in the accession data. Draw round them with the area select tool on the scattergram and see where they are located.

Weighting changes for actual use in the PCA and cluster analysis will not usually go to the extremes. Unlike the transformation, we do not have an immediate measure of how good the result is looking. For the transformation, this is the after transformation window. The scattergram is the only immediate feedback you have when altering the weights. Use this to look for interesting patterns that can then be fed through the clustering tool. If you have some prior knowledge of the germplasm on which you are working, then use that in setting the weights. If a species were known to be especially sensitive to temperature, then you would be justified in setting the temperature weight upwards. However, remember that whenever you set one or two weights very high you are losing information from the calibration set.

As you scrolled through the rainfall weights, did you notice an interesting occurrence? As you passed through weight 1.5 the scattergram coalesced into compact groups, only to fly apart again as you increased the weight. This is the sort of sign of structure we are looking for in the data. Let us set the rainfall weight to 1.5 and the other temperature weights at 0.75 each.

The Cluster Analysis

The cluster analysis function tool is included in FloraMap to investigate the possibility of multiple populations.

✓ Clear the layers menu to free screen space.

✓ Click on the cluster icon on the map window.

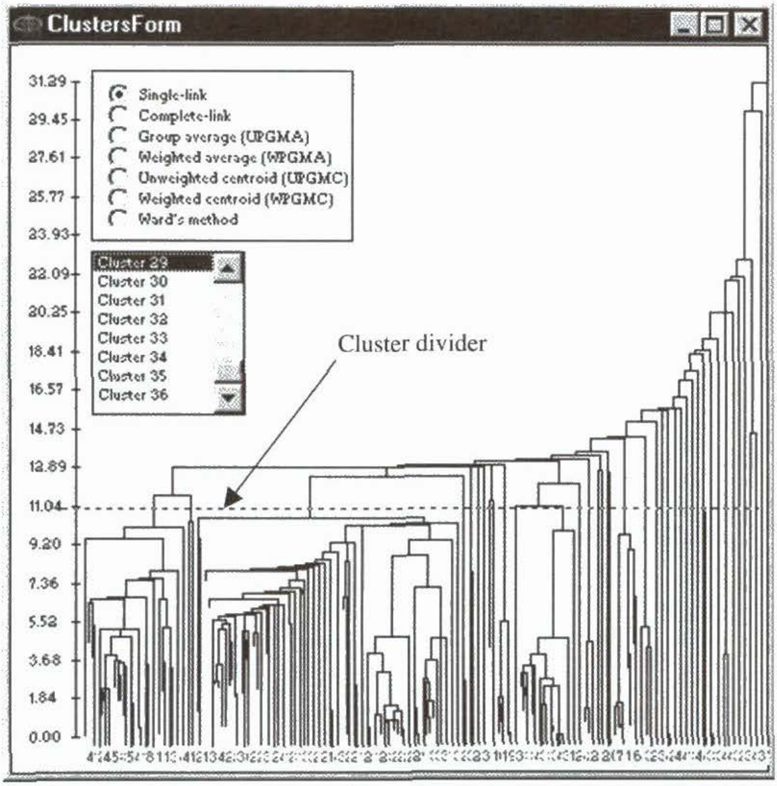
The cluster window will open and the process indicator will show calculating distances and then clustering. We have incorporated seven cluster techniques, which are all agglomerative; that is, the clusters are built up by adding members or by merging clusters. All seven use an euclidean distance measure, are hierarchic, and produce a cluster dendrogram. We apply them to the original climate dataset after transformation and weighting, but before PCA. Therefore, the choice of number of principal components *does not* affect the clustering. For more detail, see THEORY section, p 80.

No **correct** algorithm exists for clustering, and no **correct** result for a cluster analysis. Although the methods draw heavily on the mathematics of graph theory and matrix algebra, they are essentially subjective methods of finding a way to simplify complex data structures. The end result depends on what the user wants out of the analysis. In our case, we need to subdivide the data matrix into sets that will cover the various climate types that we find in the calibration set.

✓ Try out the cluster techniques. Select a method from the menu in the cluster window. The clustering will proceed automatically.

The method that you eventually select will depend on what exactly you are seeking. If you wish to simply divide the accession set into two or three clusters, then you will select a method that gives clean compact clusters, well divided, at a high level. Complete-link and Ward's method usually do this well. Ward's method is particularly suitable because it attempts to minimize the squared error term within clusters.

If you are looking for detailed structure in the dataset, then the single-link algorithm is sometimes helpful. Single-link joins clusters and adds elements by the shortest distance. It is closely related to the minimum spanning tree of graph theory. It often produces an untidy looking dendrogram because of its tendency to link members one at a time in long, strung-out clusters. It is rarely of use in providing clean cuts into major clusters, but the detail may be illuminating.



ClustersForm

● Single-link
○ Complete-link
○ Group average (UPGMA)
○ Weighted average (WPGMA)
○ Unweighted centroid (UPGMC)
○ Weighted centroid (WPGMC)
○ Ward's method

Cluster 29
Cluster 30
Cluster 31
Cluster 32
Cluster 33
Cluster 34
Cluster 35
Cluster 36

✓ Pull down the cluster divider line by clicking on the *y*-axis where the broken horizontal line crosses. Pull it down until you have 36 clusters showing.

✓ Select each cluster starting with number 36. Click right on the map surface and select view dataset from the small menu.

You will notice that the last seven clusters are single accessions or very closely matched pairs. The ragged way the single link algorithms keep tacking on the next most distant point is often considered a failing. It certainly is if what you are wanting is a clean set of a few globular type clusters.

As a tool for looking for outliers in the data it is excellent. Look at the *list of accessions* from cluster 36 down to 30. It contains:

(a) The only point that falls in Mexico,

(b) Three points with no elevation data,

(c) Our old friend Sgui2163 – tamed, but not truly brought into line at an elevation of 2150 meters, and

(d) Two pairs of points from the Chocó region of Colombia where the rainfall is exceptionally heavy.

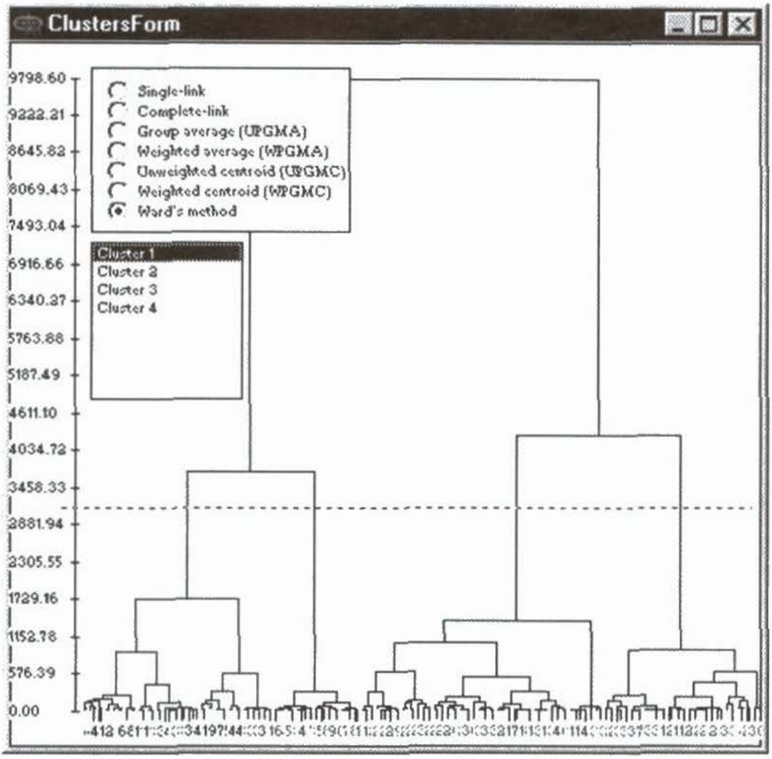
We are not going to burden you with trying to sort this out. It would take a long time and it might even mean trying to fill out some of the collections. Nevertheless, it is the sort of cleanup that should be done on a real dataset and might lead to new insights about the germplasm under study. For example, why is there only one point in Mexico and one above 2000 meters? Is this a characteristic of the germplasm or the collector?

So we will proceed with the analysis as if these outliers had not happened. Let us have another look at what you can get out of the single-link algorithm. Down on the left side of the *diagram you can* see what appear to be good cluster groupings. To reach them you will have to pull the cluster divider down until about 55 clusters are delineated.

✓ Select clusters from 1 to, for example, 16 by clicking on the cluster number in the window. Ignore the small slivers of clusters and look at the main ones. Bring up the spreadsheet with view dataset as before.

See where the clusters appear on the scattergram. The points from the selected cluster will appear in blue. Note that some have almost all accessions with the value 12 under ISO, others have predominantly the value 1. These values are the groupings from an isoenzyme study. Climate and the isoenzyme grouping appear to be related, but from this cluster analysis the relationship is not too clear.

Look now at the Ward's method clusters. These are much clearer cut and more compact. The cluster divider can easily cut the dataset into four clusters of about equal size.



ClustersForm

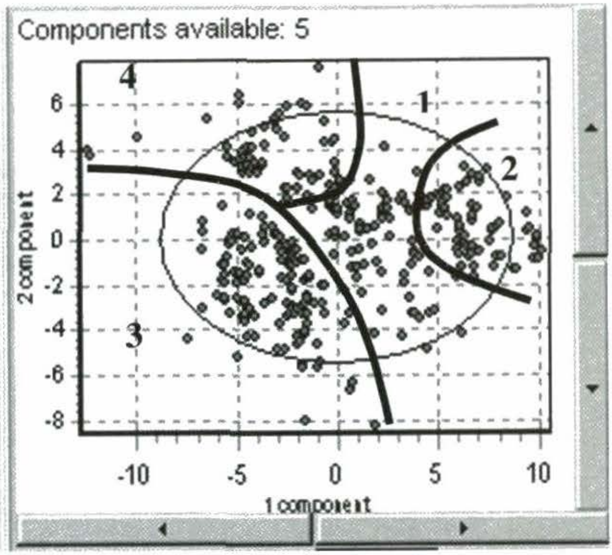
○ Single-link
○ Complete-link
○ Group average (UPGMA)
○ Weighted average (WPGMA)
○ Unweighted centroid (UPGMC)
○ Weighted centroid (WPGMC)
● Ward's method

Cluster 1
Cluster 2
Cluster 3
Cluster 4

✓ Place the cursor on a dendrogram line to view the characteristics of the cluster below that point.

✓ Place the cursor on the dendrogram, left click and draw down to the right to open a zoom window. Place the cursor on the dendrogram surface, hold down the shift key, and push the dendrogram with the mouse to pan to new areas.

Look now at the scattergram and select each of the four clusters in turn with the scattergram on the first and second component. You can see how the clusters neatly divide the scattergram into four sections. There are even traces of a space between the sections.



✓ Select each cluster in turn and bring up the spreadsheet with view dataset.

✓ Scroll through the spreadsheets noting the incidence of 12 or 1 for the ISO character.

Clusters 3 and 4 are, with only two exceptions, all ISO class 12. Cluster 2 is mixed, about half-and-half, but cluster 1 is predominantly ISO class 1. The indication that climate and the isoenzyme classification may be related is potentially important, but we will return to that later.

Let us try fitting the independent *climate* clusters as individual datasets. You can do this straight from the map window without having to exit FloraMap.

✓ Clear any probability layers from the map with the layers menu.

✓ Select cluster 1 in the cluster window.

✓ Check that the cluster is selected correctly in the scattergram.

Now you can go one of two routes towards mapping the cluster. The first serves as a quick check on what the clusters will look like on the map. The second will give you full control over the new PCA and allow you to check for data consistency.

Route 1

✓ Simply click on the map icon with the cluster selected. The PCA is recalculated for the selected cluster using the settings in the PCA window. You do not get the chance to change the options and nothing is changed in the PCA window. To see the other clusters mapped, simply select the cluster and choose the map icon again.

Route 2

✓ Click on the PCA icon. The PCA of the selected cluster will be calculated. Check that all is well on the scattergram. Sometimes, outliers appear that were not noticeable at earlier stages of the analysis. Grit your teeth and hope this does not happen, because if it does, you will have to go back to error chasing and start again. You can now apply weights, change transformation, and select number of components in the PCA as if this was a completely separate analysis (which it is). If everything is to your satisfaction, select the minimum probability for mapping. Because you will eventually be overlaying three probability surfaces, we suggest that you set it reasonably high, at 0.5 or even 0.6.

If you do find dubious points, you will want to see what information you have about them.

✓ Right click anywhere on the screen and select view dataset. A spreadsheet will appear with the accessions of the selected cluster.

☞ *The spreadsheets shown in FloraMap are minimally functional. They only exist to display the data. You cannot change, add, delete, or sort data in these spreadsheets. To do so, you must exit FloraMap and use your own spreadsheet program to make the changes.*

✓ Click on the map display icon.

✓ Click on the layer control icon and change the color to something distinctive.

✓ Clear the PCA window and spreadsheet by closing the windows. Select another cluster and click on the PCA icon to continue reevaluating the PCA and mapping the cluster.

You can save the cluster probability layers for the future using the save layer icon in the layers menu. We recommend that you rename the layer as you save it. If you leave the default names, the file may be overwritten in a later analysis. To save the selected accessions dataset hit file, save, and give it a name.

The map you have created with the four clusters from the Ward's method clustering shows a much better fit to the accessions than the single population fit, even though it does not cover all the accession points. You should not expect that the probability model would ever account for all of the points in the accession set. We assume that they are drawn from a normal population and therefore some will always fall beyond a certain probability level.

The clusters have some obvious geographic and climatic reality. Cluster 1 reflects moderate climates of coastal Brazil, Paraguay, and the Caribbean. The second cluster is mainly the hot dry climates of the Brazilian *Cerrados*, with only a very small area in El Salvador outside the region. The third includes the cooler regions of the Andes, southern Brazil, and the highlands of Central America. Finally, the fourth represents the hot wet climates of the Central Amazon, Colombian *Llanos*, and the Mosquito Coast of Central America. This is a reassuringly logical classification and mapping. Moving down the dendrogram to more clusters gives an even better tuned classification with some useful looking areas for new exploration. However, the same danger threatens in going too far with the clustering technique as we found with the number of components. The more clusters that you separate, the more degrees of freedom that you use from the data. A multitude of small clusters will fit extremely well to the points, but give you next to no predictive capacity.

☞ *We plan to incorporate some statistics of goodness-of-fit and model validity both for the principal component selection and for the clustering tool. These should be available in the next full version of FloraMap.*

We have now come to the end of this section of analysis in FloraMap using the climate data alone. You have learned what to look for in the data in terms of consistency and error. You have seen some techniques of error correcting. You have seen how to use the scattergram and how to apply the transformations and weights. You have tried out the cluster analysis to determine if more than one population is showing different climate adaptation in the calibration set. You have seen how we can look at external germplasm data from within the clusters to get an idea of how these factors are connected with the climate variability. Now we will look at applying the germplasm information directly to the mapping.

✓ Exit FloraMap and take stylo-secondfile.dbf into your spreadsheet program.

✓ Sort the file by the column ISO, and create two new accessions datasets. Name them Group_1 and Group_12 or some such.

The clustering applied to data from the isoenzyme αβ acid phosphatase now separates these accession datasets. The groups are also correlated with the varieties *Stylosanthes guianensis* var. *vulgaris* (group 1) and *S. guianensis* var. *pauciflora* (group 12). FloraMap can presently handle only one generated climate file at a time. The following procedures, however, allow you to construct a map with both the climate probabilities and the accession points from two accession sets.

✓ Reenter FloraMap and load the accessions for the first group.

✓ Proceed with the PCA and MAP operations for this dataset.

✓ Save the accession points and probability surface as map layer files, giving them distinct names.

☞ *The shapefiles are actually available in the working directory, but this feature may change in future releases.*

✓ Delete the accession points and probability layer from the map. In most circumstances you can leave the probability layer on the map, but as conflicts sometimes occur with the use of certain shapefiles it is best to clear the map.

✓ Repeat the procedure with the second accession set.

✓ Reload the map with the saved map layer files. This will load the accession points as a shapefile and so they will not go through the normal lengthy checking and climate file creation.

✓ Recolor the map to your liking and save the complete map using the save map icon.

We saw signs in some of the climate clusters that the genetic evidence and the climate adaptation might be associated. The map you have just created clearly bears this out. Each group, as defined by the isoenzyme data, is adapted to distinct climate regimes with little geographic overlap except in the southwestern Amazon.

There is one final thing we can do with these data. Suppose that you are responsible for shipping promising pasture legume germplasm to Africa. We have just shown that group 12 materials will have a range of adaptation distinct from group 1 materials. So which do you send where?

✓ Select the Africa climate grid under settings configuration general.

✓ Go to load a layer in the layers control window.

✓ Open the shapefile AFRCOUNTRIES from the \COVERAGES\ directory on the CD-ROM.

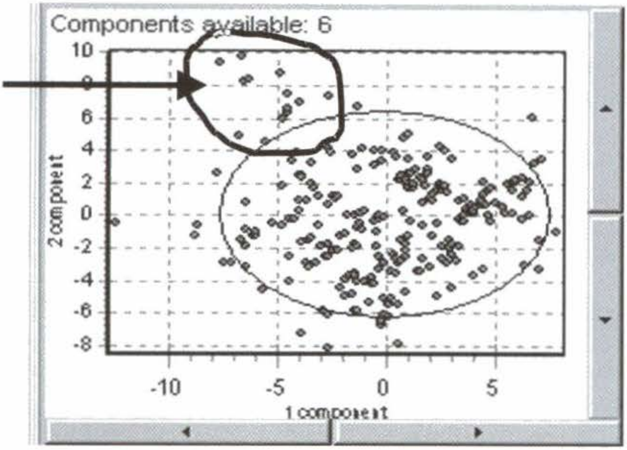
✓ Nothing will appear on the screen until you zoom out; the extra coverage for Africa has been created in the space to the right of your screen.

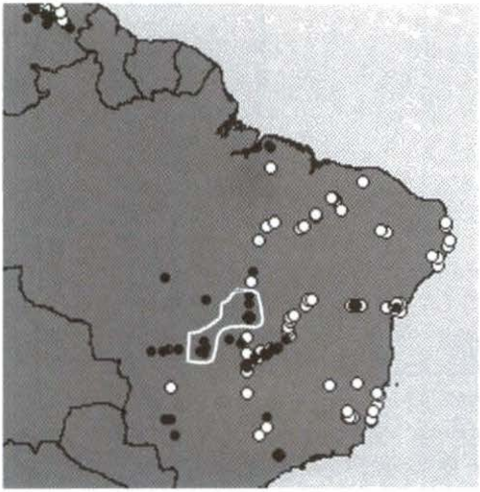
✓ Once you can see both continents, you can load the two grouped accessions files to map the potential distribution in Africa as well as Latin America.

Conclusion

If you have managed to plow through this tutorial with us you have seen practically all that FloraMap can do. On its own, it can rarely prove or disprove hypotheses. This is because of the subjective nature of many of the procedures that we employ. FloraMap can guide thought and help form hypotheses. You have seen that FloraMap can separate out populations by clustering the climates in which accessions were found. Some of these clusters can be associated with taxonomic and genetic variations. When we plot different genetic groups, they can highlight different climatic adaptation. The FloraMap system is particularly effective for interactive investigation of the datasets.

On this last note, let us look at the final analysis that you have just completed. The group 12 scattergram shows a compact group of points at the top left of the diagram that do not look like a chance result.





On the map at left, the black points are group 12 and the white, group 1. The white line encloses the points circled in the scattergram above. They are points from group 12 that are showing an atypical climate for that group and were collected in an area where the climate type is predominantly the type of group 1.

This is as far as FloraMap can take you in this particular case. The next step is to return to the germplasm and find out if this is a random occurrence or if it has some significance for these particular germplasm accessions.

3. User Reference Section

FloraMap uses a range of different windows. Some, like the mapscreen, have a top title bar showing the FloraMap application control icon at the left, the window title, and the minimize, maximize, and close icons at the right. These windows can be resized by pulling on the lower right corner. The principal components analysis (PCA) window has an application control icon, the window title, and only a close icon at the right. This window cannot be resized or minimized. The layers window has a simpler title bar and cannot be resized. When actuated it will always appear as the top level of window and should be moved to an unimportant area of the screen so that it does not obscure other screens. It can be switched off when not needed. The title bar shows dark blue when the window is active.

Because of the high demand for CPU that most of the calculations in FloraMap make, we have designed much of it to be a high-level process. This means that you will sometimes find that FloraMap windows will float in front of another application that you are using. If this annoys you, switch FloraMap out or reduce the map window to an icon while you are using an alternative application.

The Map Window

The mapscreen is the main mapping screen and is the first that appears on entering the program. The example shows a map of Latin America and consists of two layers, the basic map with the country boundaries, and the accession points for *Stylosanthes guianensis*. This pre-prepared map has been loaded from a MAP file.

Pull-down menus show along the menu bar underneath the title bar. The functions available from these are described below, and are also largely available from the service icons. The numbers shown at the bottom right of the window are the longitude and latitude of the cursor in decimal degrees; latitude is negative to the south and longitude negative to the west.

The information window will appear any time that the cursor remains unmoving on the map surface for a few seconds. It will describe the characteristics of the top layer of the map at the cursor point.

The small menu is available at any time by right clicking with the cursor anywhere on the map surface. The options it shows are:

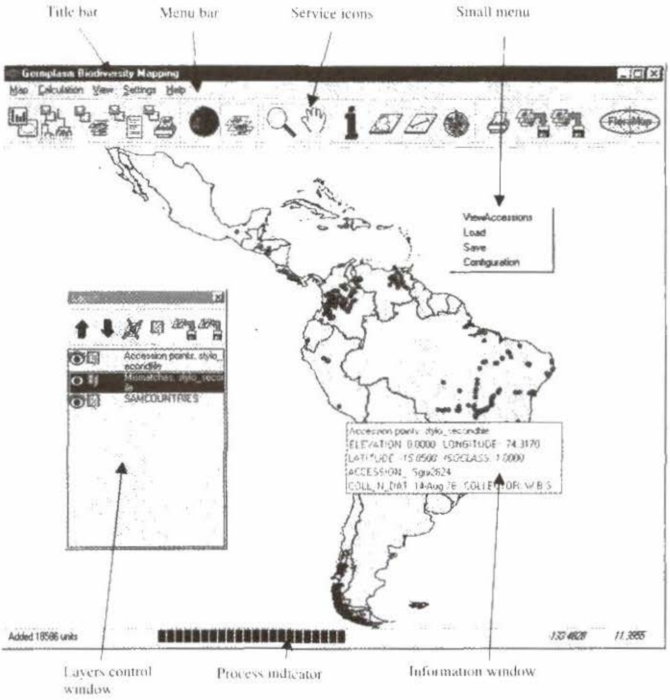
View dataset - displays a spreadsheet with the current accessions set,

Load - loads a map from a MAP file; this is equivalent to the load map icon,

Save - saves the current map with all its layers to a MAP file, and

Configuration - displays the configuration menu.

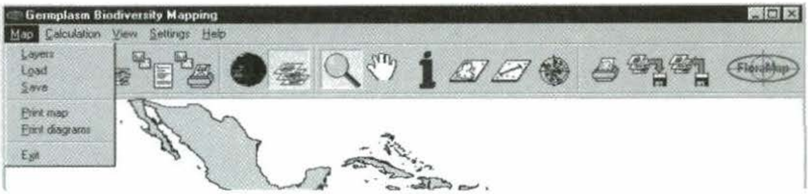
A process indicator is built into the bottom bar of the window that names the process in current operation and shows the lapsed interval.



Accession points: stylo_seconditie
ELEVATION 0.0000 LONGITUDE -74.3170
LATITUDE -15.0500 ISOCLASS 1.0000
ACCESSION_ Sgu2624
COLL_N_DAT 14-Aug-76 COLLECTOR W.B.S

Pull-down menus

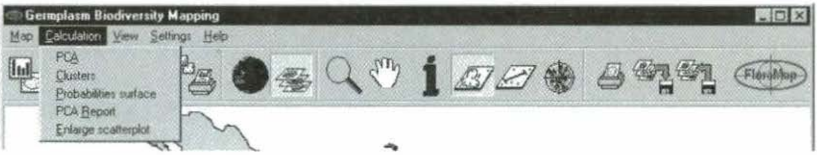
Pull-down menus include operations that are also available from the icons. More details are given later under icon explanations.



Pull down **map** and see as above

1. Layers – proceeds with the layers menu.
2. Load – loads a previously constructed map.
3. Save – saves the map you are producing.
4. Print map – prints the map.
5. Print diagrams – prints histograms of rainfall before and after transformation and all possible scattergrams.
6. Exit – exits from FloraMap session.

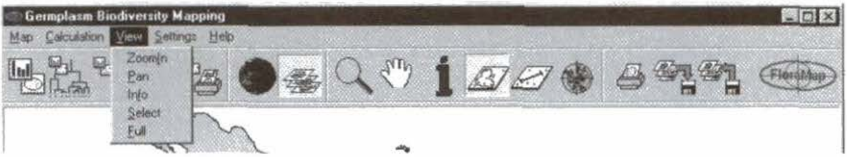
☞ *Note that all possible scattergrams can be a huge number when many components are selected.*



Pull down **calculation** and see as above

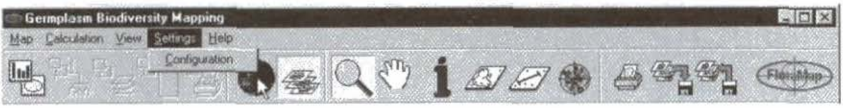
1. PCA – principal components analysis.
2. Cluster – cluster analysis.
3. Probabilities surface – calculates the probability and maps it.

4. PCA report – produces output file from PCA analysis.
5. Enlarge scatterplot – copies the scattergram window out into a larger window where individual accessions can be separated.



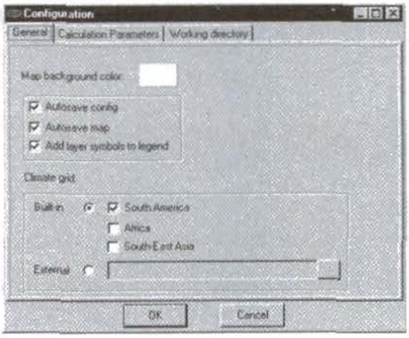
Pull down **view** and see as above

1. Zoomin – zooms in to magnify areas of the map.
2. Pan – moves the magnified map to see further details.
3. Info – shows information on the map layers.
4. Select – selects an area on the map and the points that it encompasses.
5. Full – zooms out to display the full map.



Pull down **settings** and see as above

Only configuration shows and is not on the icon list so is explained here. Click on configuration on the pull-down menu or the right click menu. The configuration window will appear. It is a scaleable window with three function windows.



Configuration

General | Calculation Parameters | Working directory

Map background color:

☑ Autosave config
☑ Autosave map
☑ Add layer symbols to legend

Climate grid

Built-in ● ☑ South America
☐ Africa
☐ South-East Asia

External ○

OK Cancel

Under **general** the following options appear.

Autosave configuration – If selected (ticked) the configuration you chose while working is kept as the default configuration when you leave FloraMap. Each time you return to the system, the last configuration that you used appears. If you made no selection, then the default configuration appears.

Autosave map gives the same option as above, but for the working map. All layers with their formats and colors are saved automatically at the end of a session as the default.map.

Add layer symbols to legend gives the option of putting a legend on the map. The option is also under control from the layers window, which will be dealt with later.

Map background color can only be set here. It will normally be a clear blue for the sea, but can be changed to other shades depending on the color printer used.

☞ *Your color printer may not faithfully reproduce the colors you see on the screen. Printers vary in the tones they can produce. Make some trial printings to determine your best background color for printing.*

Climate grids

Built-in grids

The climate grids are crucial to the working of FloraMap. You must specify which grids you will need for your analysis. There are presently three internal (built-in) climate grids that cover the bulk of the tropical regions. You can select any combination of these. The Latin American and African grids are at a pixel size of 10 arc minutes (about 18 km); the Asian grid is at 5 arc minutes (about 9 km).

☞ *The present database engine that is used in this release of FloraMap is slow with these large files.* ***Do not select climate grids unless you are definitely going to use them.*** *It is highly unlikely that you will have a calibration set that spans continents. You will only select two or more grids when you wish to map the probability surface onto two continents together on the same map.*

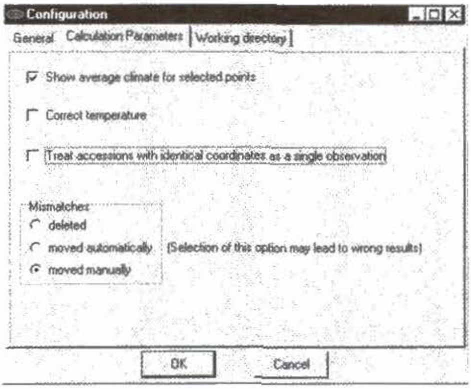
External grids

These are grid files just like the internal ones, but do not support the multiple mapping functions of the internal grids. In order to use these, all other grids must be deselected. This is done automatically when you select an external grid. The probability surface can still be mapped over another grid, but it cannot be done over two or more grids at once. External grids can be placed in any directory you wish to use.

The only external grid that is provided with FloraMap version 1 is the 5 arc minute grid of the lower 48 continental states of the USA. This is a resampling of the 2.5 arc minute climate surfaces kindly made available by Chris Daly of the University of Oregon (see Daly and Taylor 1998). You will find this in the subdirectory \external\ under climate grids in the CD-ROM.

Under **calculation parameters** you have some highly important options

Show average climate for selected points is an important new feature of this release. FloraMap will calculate the average climate for a selected group of points. This will work for points selected as a cluster on the dendrogram (see cluster window, p 65), points selected on the scattergram, or map (see area select icon, p 48).



☑ Show average climate for selected points

☐ Correct temperature

☐ Treat accessions with identical coordinates as a single observation

Mismatches

○ deleted

○ moved automatically (Selection of this option may lead to wrong results)

● moved manually

The average climate will be shown in the climate diagram (see p 52).

Correct temperature can only be used if you have elevation data in your accession dataset. We strongly recommend that you use the option if possible. You do not need to have a complete set of elevations for all accessions (see TUTORIAL section, p 18).

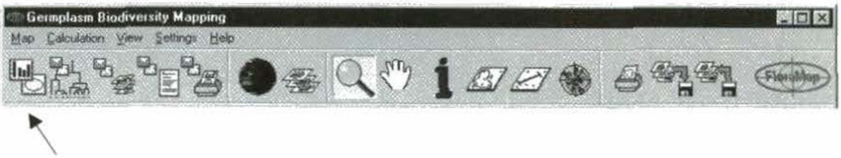
Treat accessions with identical coordinates as a single observation controls the way FloraMap interprets an accessions set. It determines whether the unit of analysis is an accession or a collection point. For further discussion, see TUTORIAL section, p 7.

Mismatches sets the way mismatches are handled. Mismatches are the accession points that do not coincide with a valid pixel in the climate grid file. (See TUTORIAL section, p 11). To set them to be deleted by default is rather drastic, but sometimes, drastic measures are needed. Setting them to moved automatically will put them on the nearest pixel in the climate grid file. This serves for points that are lying a few kilometers off the coast, but points that are further away need to be checked manually, using moved manually.

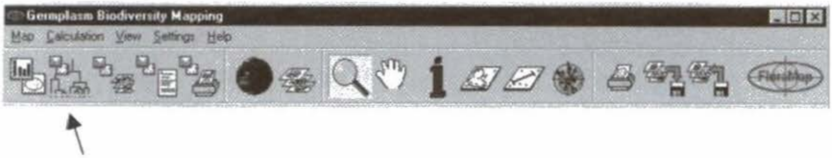
☞ *We strongly recommend that the moved automatically option is not engaged until all distant mismatched points are accounted for. Distant mismatched points are* ***almost never*** *correct when moved to the nearest point on the climate surface.*

Working directory is the working directory for your accessions datasets, probability surface, and MAP files. To work through the tutorial set it to **c:\Program Files\CIAT\FloraMap\demo\**. After working through the Tutorial you should change it to a directory of your choice. It is not recommended to carry on using a data directory under your program files directory.

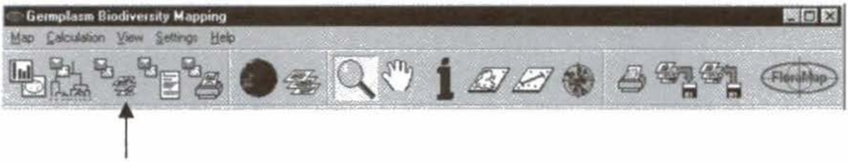
The Main Menu Service Icons



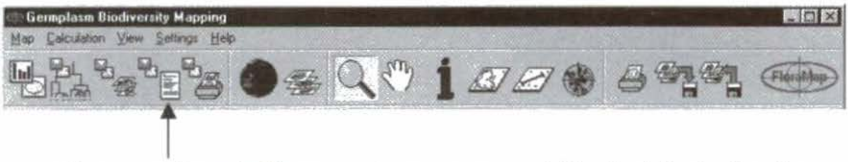
The **principal components icon** (PCA) initiates the PCA analysis with the accessions dataset that has been loaded onto the map, a set defined by the area select tool, or a selected cluster from the cluster window. For further description of the analysis see under PCA window, p 61.


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

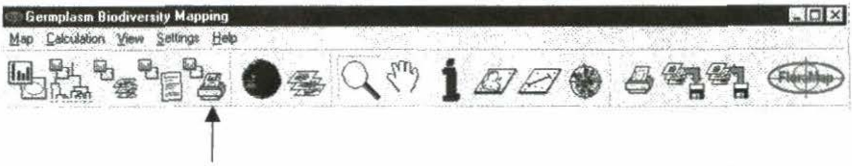
The **cluster icon** initiates clustering on the full accession set, a set defined by the area select tool, or any set defined in the PCA window. The last can be used to produce a form of recursive clustering. The cluster icon cannot initiate clustering on a selected cluster, but if the selected cluster is introduced into the PCA by the principal components icon then the cluster icon will use that as the set for further clustering. It is important not to set clustering until required, because operations in the PCA window will cause automatic clustering. This can lead to slow operation if the accession sets are large.


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **map display icon** takes the probability model currently calculated in the PCA window and plots the probability layer on the map in the map window. Calculating the probability surface can take some time. If you have initiated this and change your mind, it can be stopped by toggling the icon again.

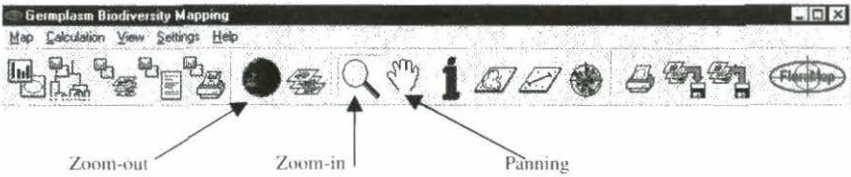

Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **report model icon** produces a report file that includes the raw data, transformation details, means and variances, and transformed data. The eigenvalues and component loadings (eigenvectors) follow these. Finally, it contains the component scores for the number of components selected. When the calibration set is large and many components are selected, this can be a very large file. We recommend that you inspect it before printing.


Germplasm Biodiversity Mapping
Map Calculation View Settings Help

The **print diagrams icon** prints the diagrams from the PCA window. It will print all 12 monthly rainfall histograms before and after transformation, and all possible scattergrams.

☞ *Note that all possible scattergrams will be a large number when many components are selected.*


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
Zoom-out
Zoom-in
Panning

The **zoom icons** are three related icons. We will deal with them together.

Try the zoom-in icon

- ✓ Place the cursor on the map and pull out a rectangle over the area you want amplified. FloraMap will adjust the longest axis of this rectangle to fill the map window.

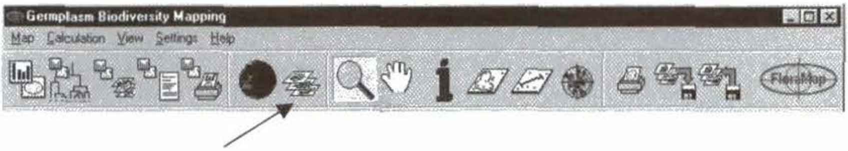
Once you have zoomed in to an area, you may need to move around.

- ✓ Toggle the panning icon and place the cursor on the map.
- ✓ Click left and push the map around with the cursor. Clear space will appear in the area that you have pushed it away from, but it will quickly fill up with new map coverage as the pan proceeds.
- ✓ Let go and grab the map again by releasing and clicking the left mouse button.

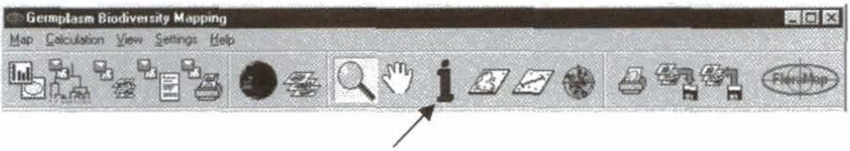
☞ *You can release and move the map faster than the screen can refill so take care not to get lost. FloraMap will remember all the movements you have made and keep moving the map until all your pans are fulfilled.*

Zooming out is simple

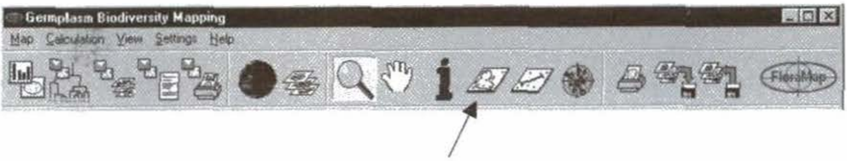
✓ Hit the zoom-out icon. The complete map will be redrawn. Future versions may have a more sophisticated zoom-out facility, but for the present you go back to start, do not pass go, and do not collect $200.


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **layers icon** controls the map layers. Use this icon to switch the layers control window on and off. With the window switched on you have individual control of all of the layers in the map. See p 56 for detailed description of the layers control window.


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **information icon** is an active method of finding out information on any part of the map. Select the information icon and click the cursor on the required position. The map element will flash momentarily and a small spreadsheet will appear. This may be moved to another part of the screen and will remain until you close it, even if you delete the map layer to which it refers. Selecting another icon from the menu deactivates the information icon.


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **area select icon** is used to identify sets of accession points as they appear in the map window and in the scattergram in the PCA window. See p 61 for the window description.

Unlike some of the other icons that open windows, the area select icon operates directly on both the PCA and the MAP windows. It is one of the most useful tools provided as an icon. As you will see from the TUTORIAL and the THEORY sections, it is often necessary to consider the calibration set of accessions as a number of different populations. The area select icon is in many cases a highly flexible way of doing this. You can use this icon to identify subsets of the accessions set, separate them out, and store them as sets for separate analysis, or even proceed directly to a new analysis.

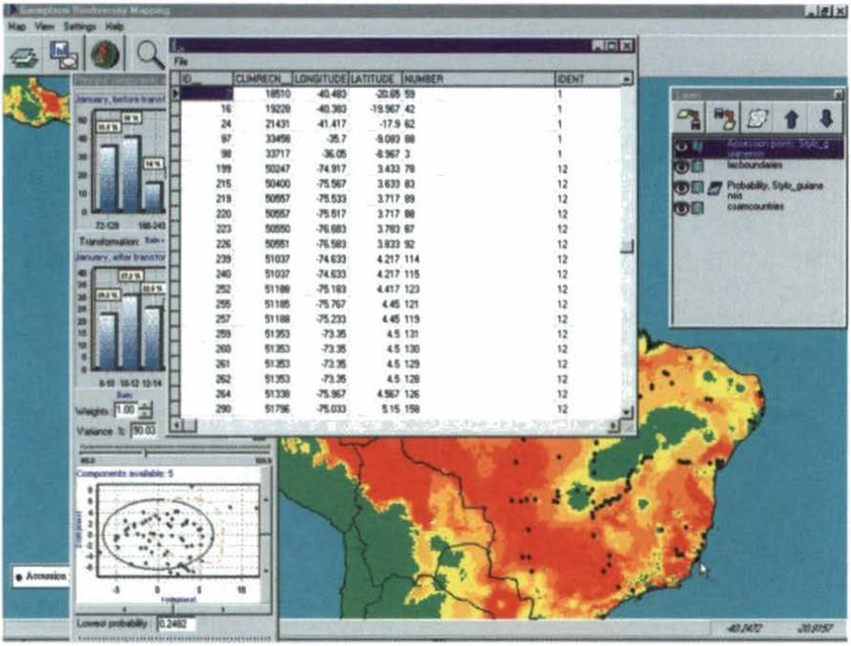
To use the tool, draw around the points on the map or in the scattergram window. A flashing red line will follow the cursor when the left button is held down on the mouse. The polygon will close automatically as the button is released. If you are drawing around points on the map and you have more than one set of accession points loaded you must select the relevant set in the layers control window. Thus, if you want to identify a set of mismatched points, the mismatched set must be highlighted.

The area select tool is always available on the enlarged scattergram window regardless of whether or not you have toggled the icon.

If you have set the option show average climate for selected points in the configuration menu, the average climate will be calculated and displayed in the CLIMATE DIAGRAM (see p 52). The average is calculated on the rotated climate data to ensure that the climate records match for the calculation. Once it is shown in the climate diagram, you will notice that there is no **rotate/unrotate** option as there is when looking at the climate of a single accession or point from the map. This is because rotation has no meaning when applied to an average climate. The records may have come from geographically diverse points and so there is no correct phase angle for rotating back the average.

☞ *If you draw around one single point, you will still get no opportunity to unrotate the data. In this case, simply point at the accession with the climate diagram tool.*

Below, we have toggled the area select icon and drawn a freehand polygon to enclose a set of accessions as they appear on the scattergram.



ID	CLIMRECN	LONGITUDE	LATITUDE	NUMBER	IDENT
7	18510	-40.483	-20.65	59	1
16	19229	-40.383	-19.967	42	1
24	21431	-41.417	-17.9	62	1
97	33458	-35.7	-9.083	88	1
98	33717	-36.05	-8.967	3	1
199	50247	-74.917	3.433	78	12
215	50400	-75.567	3.633	83	12
219	50557	-75.533	3.717	89	12
220	50557	-75.617	3.717	88	12
223	50550	-76.683	3.783	87	12
226	50551	-76.583	3.833	92	12
239	51037	-74.633	4.217	114	12
240	51037	-74.633	4.217	115	12
252	51188	-75.183	4.417	123	12
255	51185	-75.767	4.45	121	12
257	51188	-75.233	4.45	119	12
259	51353	-73.35	4.5	131	12
260	51353	-73.35	4.5	130	12
261	51353	-73.35	4.5	129	12
262	51353	-73.35	4.5	128	12
264	51338	-75.967	4.567	126	12
290	51796	-75.033	5.15	158	12

The freehand encircling polygon appears as a dashed red line that revolves around the points to draw attention. The selected points have changed color from green to dark blue, and the spreadsheet window has appeared on the screen. The spreadsheet now holds the dataset of the selected points. The accession point with ID 7 is selected and the point flashes on the Brazilian coast just north of Rio de Janeiro where the cursor is pointing, but paper has no response time, so you cannot see the flashes here on the page.

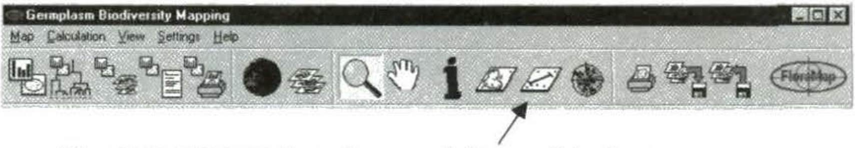
Scroll through the scattergrams and follow where the selected points fall. This is an ideal way of getting a good idea of the distribution of the accession points in a multidimensional space.

With the area select icon toggled you can do exactly the same for any group of points on the map window. Click the cursor on the map and draw around a set of points. Notice that the PCA analysis is changing constantly. The points in the scattergram change color as you change the dataset. It is continually updated and completely interactive. If you like the look of a subset you have chosen, you can return to the PCA window and change the transformation or the weightings, the probability setting, or number of components.

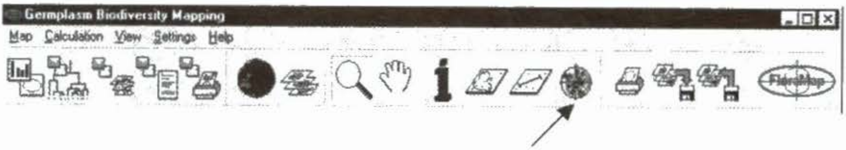
Filing the spreadsheet is standard. Hit file, save, and give it a name. You can then drop out of the analysis you are in at present, and restart using the new subset. Or, if you do not have time for that, hit the PCA icon and the PCA analysis will be done on the subset you have designated. You can then alter the model parameters (transforms, weights, and number of components). If you have even less time and more impatience, hit the map display icon and the probability surface for the model of the chosen subset will be calculated and displayed on the map. In this case, you will have to accept the model parameters as they were set on the previous analysis.

It is easy to get lost in the complexities of the different modes of selection. You can subset on the map, return to the scattergram, which now contains only that subset, and subset again. We recommend that, unless you are just having fun to see how far you can push our product without breaking it, you keep careful notes of what you are doing and file the spreadsheet regularly under names that you can follow in the analysis. If you choose the fast routes to the analysis and display, they will be assigned names by adding a sequence number to the end of the file name.

It may be that you will want to draw round a number of subsets that you would like to analyze together. To do this, select and file the spreadsheets separately and then concatenate them in your spreadsheet program.

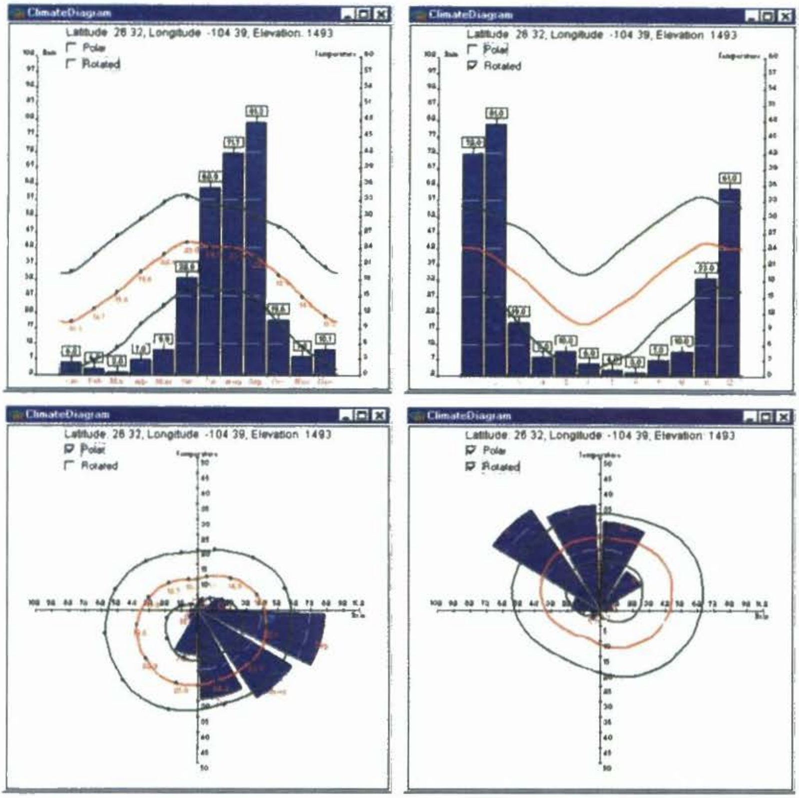

Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **map distance icon** is especially useful when you are debugging an accessions dataset. Toggle the icon, place the cursor on a point on the map, click left and drag the line out to measure a map distance in kilometers. The distance is given in the lower left corner of the map window. With this you can measure the distance between doubtful points or the distance of an accession point to geographical features such as roads (see p 20).


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

The **climate diagram icon** is a useful tool for investigating the structure of an accessions set, for debugging the accessions, and even for just browsing the climate database. Toggle the icon and point at either an accession point or any point on the map. FloraMap will search for the relevant climate record (it is a little slow in this release – we promise to speed it up for the next).

You can toggle the climate diagram between any of the following four representations.



ClimateDiagram
Latitude 28 32, Longitude -104 39, Elevation 1493
☐ Polar
☐ Rotated

ClimateDiagram
Latitude 28 32, Longitude -104 39, Elevation 1493
☐ Polar
☑ Rotated

ClimateDiagram
Latitude 28 32, Longitude -104 39, Elevation 1493
☑ Polar
☐ Rotated

ClimateDiagram
Latitude 28 32, Longitude -104 39, Elevation 1493
☑ Polar
☑ Rotated

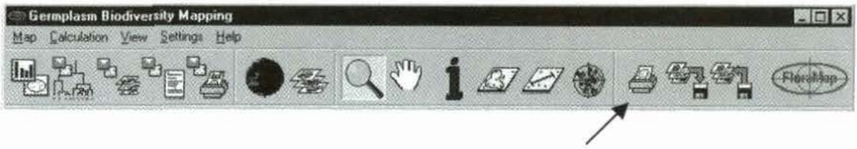
The two diagrams on the left show the monthly rainfall totals, the mean, maximum, and minimum temperatures in Cartesian and polar coordinates. The polar plot starts at the top and runs anticlockwise from January to December. The rainfall totals label each column of the histogram, and the monthly mean temperatures are shown against the mean temperature plot.

The two diagrams on the right show the rotated data. This is not a simple monthly shift (see THEORY section, p 69), but is produced by a special Fourier transform that conserves the totals of each 30-day period and the overall yearly mean. Because the rotation angle can be a fraction of a month, the plotted monthly values change, as you will notice in this example from the central highlands of Mexico. You can readily see the rotation that shifts the June to October rainy season to the start of the virtual rotated year.

There are no months marked on the rotated year because the zero point of the time scale is arbitrary.

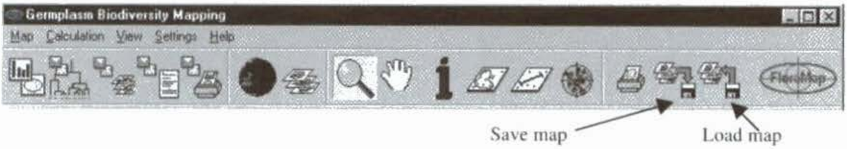
If you have set the option show average climate for selected points, a climate diagram will appear automatically when you select points with the area select tool or on the cluster diagram. In both these cases, the rotation option will not appear. This is because the average climates have been calculated using the virtual zero points of the rotated data. The average rotation angle has no meaning and so there is no angle through which to unrotate the record. It exists perpetually in a virtual rotated space.

☞ *There is no facility for printing the climate diagram, but if you hit alt-print screen while the window is active (i.e., it has the title bar in blue) the window will be copied into the edit clipboard. You can then paste it into a Word document, spreadsheet et cetera.*


Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap

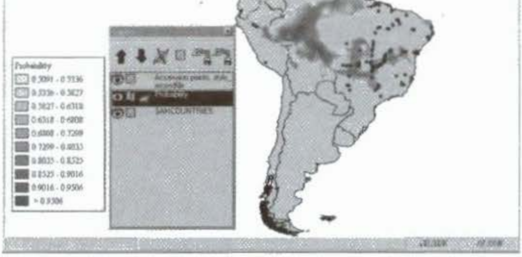
The **print map icon** sends the map, as shown in the map window, to a printer of your choice among those that you have available. It will not print the screen surrounds or any overlapping windows. The map consists of the active layers in the layers window plus any legend information that is selected.

☞ *Different printers handle colors differently. They are almost never as they appear on the screen. Be prepared to experiment and get to know the colors from the menu, or from your custom colors menu, that print well on the printers you are using.*

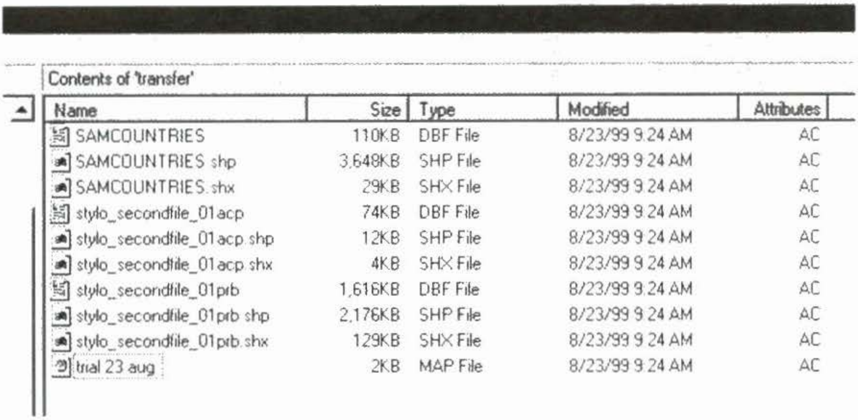

Germplasm Biodiversity Mapping
Map Calculation View Settings Help
FloraMap
Save map
Load map

The **save map icon** does as the icon picture illustrates—it takes all the layers that you have active in the layers window and places references to them in a MAP file in the working directory. The MAP file is a metafile (see APPENDIX A). It does not contain any layer data itself but contains the information on which layers are displayed with which display options in order to recreate the map. Each layer of the map is saved separately as a shapefile in the directory of your choice. Each shapefile consists of three parts, the **DBF** file, the **SHP** file, and the index file **SHX**.

Here is an example. The map below has the three layers – stylo_secondfile, probability, and SAMCOUNTRIES. The map has



been saved into the directory **\transfer\\.** The result is the set of 10 files shown below.



Contents of 'transfer'

Name	Size	Type	Modified	Attributes
SAMCOUNTRIES	110KB	DBF File	8/23/99 9:24 AM	AC
SAMCOUNTRIES.shp	3,648KB	SHP File	8/23/99 9:24 AM	AC
SAMCOUNTRIES.shx	29KB	SHX File	8/23/99 9:24 AM	AC
stylo_secondfile_01acp	74KB	DBF File	8/23/99 9:24 AM	AC
stylo_secondfile_01acp.shp	12KB	SHP File	8/23/99 9:24 AM	AC
stylo_secondfile_01acp.shx	4KB	SHX File	8/23/99 9:24 AM	AC
stylo_secondfile_01prb	1,616KB	DBF File	8/23/99 9:24 AM	AC
stylo_secondfile_01prb.shp	2,176KB	SHP File	8/23/99 9:24 AM	AC
stylo_secondfile_01prb.shx	129KB	SHX File	8/23/99 9:24 AM	AC
trial 23 aug	2KB	MAP File	8/23/99 9:24 AM	AC

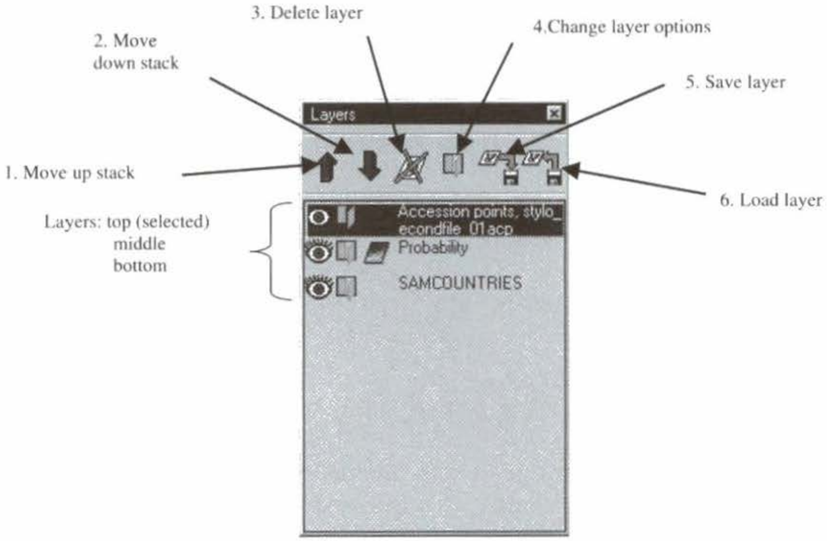
The MAP file 'trial 23 aug.map' is accompanied by the three shapefiles, each consisting of three components.

You cannot use the **load map icon** until a .MAP file has been created. One .MAP file comes with the system, the default.map. If you have set the option autosave map in the configuration window, the map that you have at the time you exit FloraMap will be saved in this MAP file.

☞ ***IMPORTANT!*** *The map that you load from a MAP file consists of only the shapefiles that create the image. The accession set that you see on the map (if you have one) is only the graphic representation. If you try to calculate a* ***PCA*** *from this map, the system will ask for an accession set to be loaded. This is because no climate file is associated with the map. To proceed with the analysis, delete the accession points layer from the map and select the accession points file once more. FloraMap will then proceed to build the climate file.*

The Layers Control Window

This is a layers control menu. The one below specifies three layers. There can be as many as required, but overindulgence clutters the map.



Icon 1 moves the selected layer up the stack.

Icon 2 moves the selected layer down the stack.

Icon 3 deletes the selected layer from the map, but not from the directory.

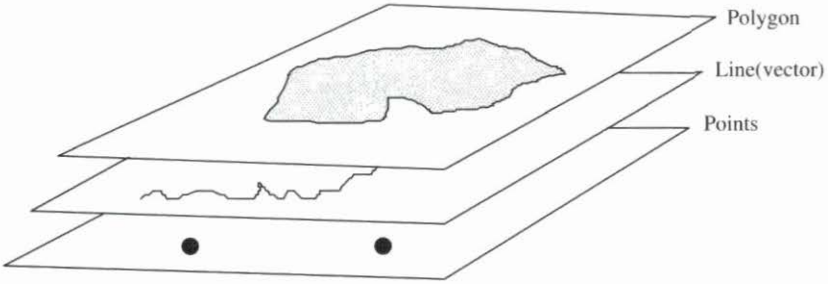
Icon 4 changes the options for the selected layer.

Icon 5 saves the layer.

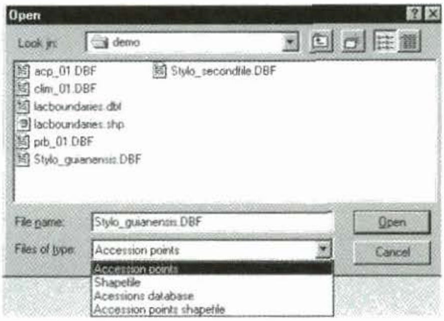
Icon 6 loads a layer.

The stack controls what is seen on the map and also its display. The layers are placed on the map from bottom to top. Upper layers obscure lower layers. The order of the layers can be changed using the arrow icons or by clicking and dragging the layer into the place required in the stack.

Polygon layers with color fill will obscure line or point data below them (see following diagram). Make sure that your map is correct by ordering the layers. A new probability layer always appears as the top layer. If you have not got the layers control window showing, click on the layers icon and reorder the accession points so that you can see them.



Loading a layer with the load layer icon covers both the loading of passive map layers (such as background, country boundaries, rivers, roads, and cities) and the active layer, which consists of the accession points set to be analyzed. The menu display at right will show the working directory where the layers are stored and, at the bottom, the types of files that you can load.



Look in: demo

acp_01 DBF
clim_01.DBF
lacboundaries.dbf
lacboundaries.shp
prb_01 DBF
Stylo_guianensis.DBF
Stylo_secondfile.DBF

File name: Stylo_guianensis.DBF

Files of type: Accession points

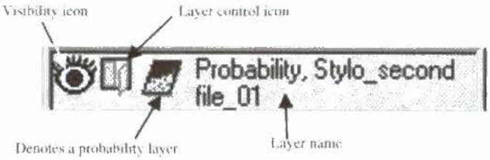
Accession points
Shapefile
Acessions database
Accession points shapefile

An accession points file can have a .dbf or .acp extension. The DBF files are in dBASE 4 format. The ACP files are space-delimited ASCII files with column headings. For further information on FloraMap files see APPENDIX A.

Shapefiles are compounds of three files with extensions – .shp, .shx, and .dbf. They are ESRI standard files and are compatible with ArcView.

☞ ***EXTREMELY IMPORTANT!*** *FloraMap has no way to differentiate an accession points file in DBF format from the DBF file of a shapefile or a climate file. If the file contains columns for latitude and longitude the file can be loaded as an accession points file. Because many shapefiles are very large, this can be a major error.*

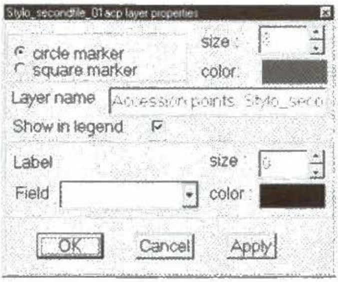
Customizing map layers


Visibility icon
Layer control icon
Probability, Stylo_second file_01
Denotes a probability layer
Layer name

The **visibility icon** is useful when composing a map. It suppresses or shows a layer in response to a double click. The layer stays logically connected to the map and so does not have to be erased and reloaded.

Double click on the open window **layer control icon** on the selected layer and you open a window to set the characteristics of that layer on the map. The first example is the accession points layer. You can set the type of marker and its color and size. Click on the color shown and you will open the color menu. Either choose from the offered palette, or mix yourself a custom color. Remember that colors from a color printer do not match what you see on the screen, so you will need some practice with this.

The **layer name** is the one that appears on the legend. The name is set by default from the filename of that layer, but it can be changed on the map to whatever wording you want for the legend. If you take off the check mark in show in legend, the layer will not be referred to in the legend.


Stylo_secondfile_01acp layer properties

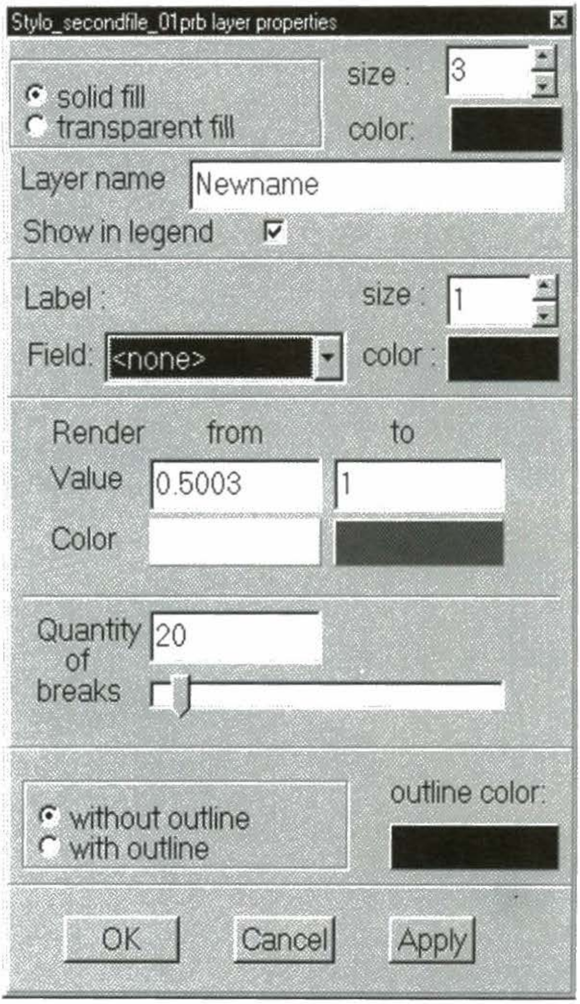
● circle marker
○ square marker
size :
color:
Layer name Accession points Stylo_seco
Show in legend ☑
Label size : 6
Field color :

Using the label fields can sometimes be useful for checking individual elements that you have on the map, but it is usually too cluttered for labeling an output map. The system will label each separate feature in the shapefile. This may function for sparsely plotted accession points. However, showing the labels on a country shapefile, labels not just the country, as you might think, but also every small island, inlet, and lake.

> ☞ *Unfortunately, the label fields scale with the map as it is zoomed; as you zoom in, they do not get less cluttered, just bigger. We will try to resolve this problem in future releases.*

Once you have produced a probability surface, FloraMap will plot it over whatever map you have in the map window. Toggle the layers up and down using the arrows in the layers window, or push them into place with the hand icon on the cursor.

A probability surface is different to the background surface that we have seen before. It is many-valued. In fact, each pixel can take any value from the minimum that you specified on the PCA screen to the value 1.0. The surface is therefore colored with a range of colors to denote the range of probabilities. You can change the colors and number of breaks in the range.



Stylo_secondfile_01prb layer properties

● solid fill
○ transparent fill

size : 3

color: ■

Layer name Newname

Show in legend ☑

Label :

size : 1

Field: <none>

color : ■

Render	from	to
Value	0.5003	1
Color	□	■

Quantity of breaks 20

● without outline
○ with outline

outline color: ■

For a probability surface the size option has no meaning. If the fill is changed to transparent it will not be seen. The label field should not be set as it has a disastrous effect on the map.

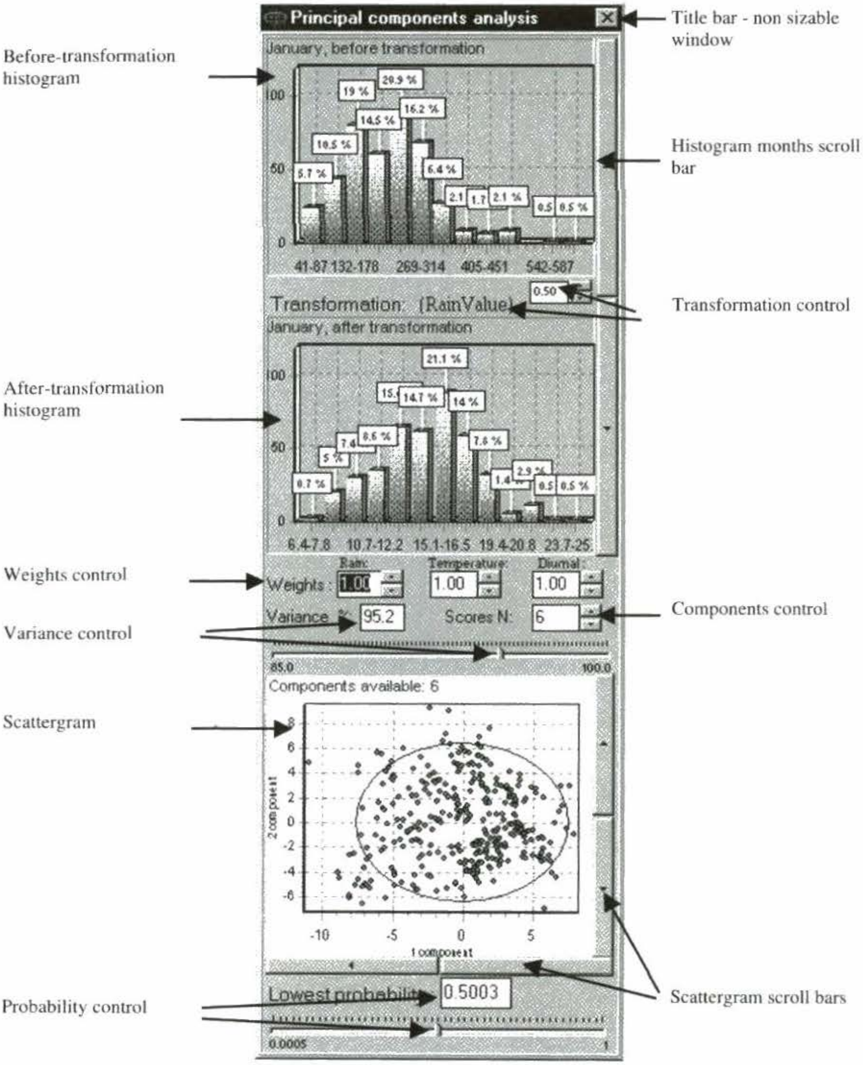
The value fields, render from and to, are automatically set to the lowest probability that you requested with the probability control on the PCA window. Very rarely will these fields need to be reset.

The hardware that you are using may constrain the number of breaks you choose. A 256-color screen-driver limits the number of breaks that are unique. An artistic limitation is at present an unfortunate restriction in the system. A color sequence cannot be defined to go from one color through a second to a third color. This would be a highly useful technique to show the detail of the probability range. For instance, the sequence black, brown, red, yellow, white cannot yet be defined, although it is a striking natural sequence to the eye. We will be working to include this flexibility in the future. At present, you can specify the end points and the color sequence assigned will be the simplest way to get from the first to the second color.

☞ ***NOTE:*** *The range break contours that you see on the map on the screen are not real lines. The pixels in the surface continue to maintain their original values. This is unlike some GIS systems that require the image to be reclassed to fix the range of classes. A consequence of this is that the range boundaries do not behave like polygons. If you switch on outline on this menu you will get the outline of each pixel, not the probability level polygons. If you wish to vectorize these, you will have to export the layer to another system that can handle raster-to-vector conversion.*

The Principal Components Analysis (PCA) Window

This window cannot be resized. If a piece is missing at the top or the bottom, then your screen does not meet the minimum configuration for the application (see GETTING STARTED).



Principal components analysis

January, before transformation

Transformation: (RainValue) 0.50

January, after transformation

Weights: Rain: 1.00 Temperature: 1.00 Diurnal: 1.00

Variance %: 95.2 Scores N: 6

Components available: 6

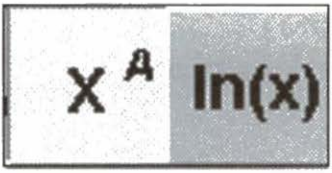
Lowest probability: 0.5003

The two histograms show us the rainfall data for the accession points dataset. These are the frequency diagrams for each month before and after transformation. Use the histogram month scroll bar at the right to select the month. Click on the histogram and pull down to the right to create a zoom window with the cursor.

Click anywhere on the histogram image and lock the cursor onto it by clicking right. Now you can pan with the mouse to scroll to the detail you want in whichever direction you want. To return to the original image, click on the histogram with the left mouse button, creating a zoom window, and pull up and to the left, the opposite of what you did to begin with.

The transformation is critical to the analysis. The whole analysis hinges on the fact that the variates are distributed as NORMAL variates. Temperatures are usually distributed normally with a beautifully symmetrical bell shaped curve. Rainfall almost never is. In fact, most studies treat it as a gamma distribution. You can see its typical shape in the January figures in this example. It is skewed heavily towards the smaller values. We cannot use a gamma distribution in FloraMap because of the probability calculations we will be making later. Therefore we must transform it to be as normal as possible.

Click anywhere on the transformation control between the two histograms and you will be given the option of transformations as at right. The two basic transformations are the natural logarithm and the power transform. Choose between the two by clicking on the menu. If your choice was the natural logarithm, there is nothing else to do.



If you chose the power transform, then you will have to specify the power in the small window provided. The range allowed is from +3.0 to +0.1, then -0.1 to -3.0. This gives you a wide range of powers from which to choose. It runs from *x* cubed to the reciprocal of *x* cubed. It skips over the range +0.1 to -0.1 because in this range the data transforms almost to the fixed value 1.0. Much of this range produces rather odd transformations, but it is useful to have the option. A good transformation to try as a start is the square root, or *x* to the power 0.5. Choose which gives the best-shaped bell-curve in the after-transformation window. We have to use the same transformation for all 12 months, so please go carefully through each month and pick the transformation that does the best job on the months with the worst distributions. Rarely will you find one that is universally good so the best compromise must be used.

You can alter the weightings that are applied to the climate variates with the weights control. Toggle the buttons and watch how the weights change. All weights have to add to 3 so they will all change as you alter one of them. The algorithm for assigning the change to the remaining weights is not always predictable. If you experience difficulty getting the weight you need, enter the numbers directly into the windows.

Watch how the scattergram, below on the screen, changes as you change the weights. We will go into the statistical implications of that later. For now, how fast does it move? It is recalculating the principal components analysis as you change the weights. How fast it can do this depends on your computer (see GETTING STARTED).

The number of component scores is an important consideration for your analysis (see THEORY section, p 72-73). The components are in descending order of variance and usually the first few of them account for a relatively large proportion of the overall variance, but **how** few, and **what** proportion?

Pick up the toggle of the variance control with the cursor and move it. The variance in the window above changes, as does the number in the components control. This is showing you the variance accounted for by the number of component scores you have selected. An alternative way is to toggle up or down on the buttons by the components control window, or to set the numbers yourself in the windows.

The scattergram is a highly important part of the analysis (see TUTORIAL and also THEORY section). There are N.(N-1)/2 diagrams when you have N scores. Thus there are three with 3 scores, six with 4 scores, and 45 with 10 scores. They all give you information. The ellipse is set at two standard deviations from the origin of the two components shown. You can toggle between scattergrams by using the four scattergram control bars. The bottom two will move you up and down the component shown on the *x* scale, the two on the right will move between components on the *y* scale.

Enlarging the scattergram

The scattergram is such a useful tool for investigating the properties of sets of accession points that we have provided the option of enlarging the window in order to see the points more clearly. Pull down the calculation menu and click on enlarge scattergram. Once the scattergram is duplicated into the large window the area select tool is automatically selected for it, even though the map window behind it becomes locked out. You can choose sets of points and scroll through the principal components on the enlarged window without affecting the one in the PCA window. Selecting a set of points will show the spreadsheet of those points. If you have the climate diagram option set on you will see the average climate for the selected set displayed.

When you have decided on a set of points to investigate further you should check OK and the set will become the selected set in the PCA window and you can proceed to enter them into a further pass of analysis. You will notice that they become blue on selection, but if you have scrolled the enlarged window to a different combination of components then you will have to scroll the small scattergram window to match. If you close the large scattergram with the Windows X button or the cancel button, nothing will be changed.

The probability control sets the lowest probability that you want to be mapped. This is important when you are constructing a map, especially if you want to show more than one probability surface, because the probability surface is a densely colored layer and other features do not show through it. The lower the probability you choose to be mapped, the more completely you will tile the map with the surface. Choose a value by typing in the box or by moving the control button left or right.

The Cluster Window

We saw above how we could manually subset the data either on the map itself or on the scattergram window. Good reasons often exist for thinking that the climates of the accession set are not homogeneous (see THEORY and TUTORIAL sections). Drawing around groups on the scattergram using the area select icon is often as good a way as any of subsetting the data. If you can see the groupings, you can delineate relatively complex clusters where a numerical algorithm may fail. However, you are restricted to looking at two dimensions at a time. Indeed, although we can visualize three-dimensional space, few people can visualize four dimensions.

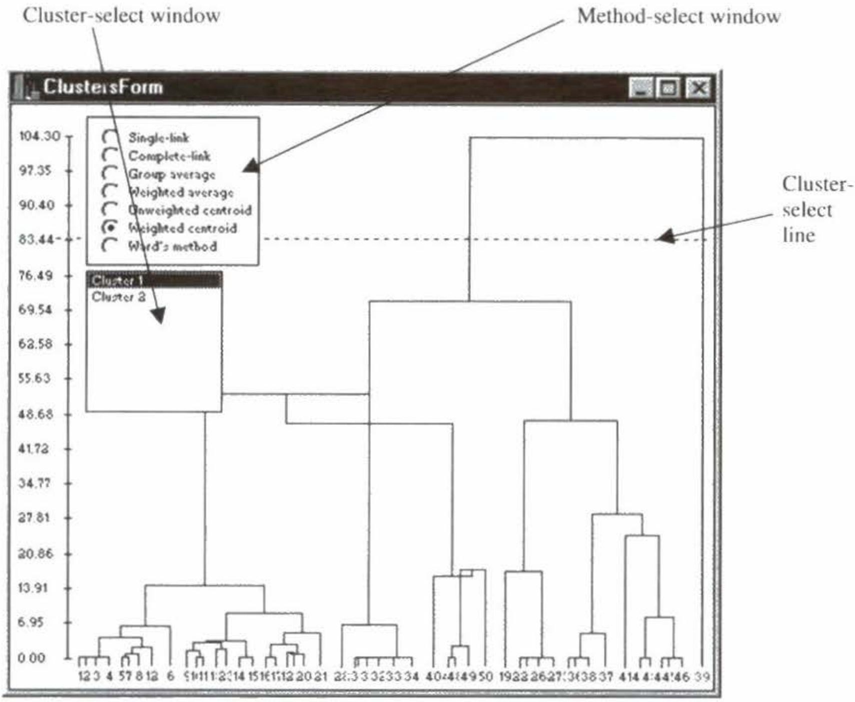
Numerical cluster analysis is a powerful (but highly subjective) tool. We have incorporated a range of cluster analysis methods to help with this problem (see cluster analysis, THEORY section).

To turn on the cluster window, select it from the calculations pull-down menu, or toggle the cluster icon.

The cluster display is fully interactive with the PCA. As you adjust the PCA, the cluster window changes to the clustering.

The heart of the display is a dendrogram showing the relationships between accessions. The scale at the left is the cluster distance; the bottom line is not a scale in the usual sense, it is showing all we can show of the accessions' identities. In most cases, the accession set will be so large that these identifiers are too crowded to show on the full diagram. Therefore, we have implemented zooming on the diagram.

Place the cursor on the diagram and pull down and to the right while holding down the left mouse button. The enclosed part of the dendrogram will expand to fill the frame. It will normally only be useful to zoom onto a complete or part cluster with the accession identifiers visible at the base of the dendrogram. However, you can zoom into any area to look at the cluster links if you wish. To zoom out, place the cursor on the diagram and move left and upwards while holding down the left button. The diagram will zoom out to its full size. The zoom function is purely to manipulate the display. It does not select accessions for further processing. With the display zoomed, you can pan on the dendrogram area by holding down the shift key while moving the cursor across the dendrogram area.



ClustersForm

- ○ Single-link
- ○ Complete-link
- ○ Group average
- ○ Weighted average
- ○ Unweighted centroid
- ● Weighted centroid
- ○ Ward's method

Cluster 1
Cluster 2

Two non-scaleable menu windows are fixed to the display area. The method-select window is the top one, which selects the cluster method. Seven of these are available at present (see THEORY section, p 80); only one can be selected at a time. The display will be recalculated as you select each different method. Try the various methods and watch the results for your dataset. There is no correct method; each has its advantages for different datasets.

Selecting, analyzing, and saving clusters

The cluster-select window is the lower one, which shows the number of clusters you have selected. To select cluster levels, click left on the left-hand cluster distance axis. This will bring the horizontally dashed line, the cluster-select line, to the cursor arrow tip. Move the cursor up and down the screen to select the cluster level you want. The number of vertical connections cut by the dashed line determines the number of clusters shown in the window. Click on a cluster within the window to highlight and select it. You can only select one cluster at a time. The selected cluster will be highlighted in blue on the dendrogram. All other cluster connections will be shown in black, and the accession identifiers are shown in red. At the same time, watch the scattergram. The points for the selected cluster will change to blue. You can scroll through the scattergrams to determine which of the components are affecting the clustering.

☞ ***NOTE:*** *The clustering is calculated from the scaled transformed variates. The principal components are another way of visualizing the variates, but each scattergram represents only the variance accounted for by the two components shown.*

You now have three options for further analysis of the selected cluster.

✓ Click on the map display icon and the probability surface calculated from the cluster points will be displayed on the map. A new PCA is calculated for the cluster of points, but the PCA display window is not changed.

✓ Click on the principal components icon and the PCA will be recalculated. The histograms and scattergram will be changed to reflect the new PCA of the cluster points. The probability surface will not be displayed until you use the map display icon again.

✓ Click right on the mouse with the cursor anywhere on the map and the small menu appears. Click on view dataset and a spreadsheet appears with the accessions data for the selected cluster. This can be named and saved in the normal way. If you wish to proceed to PCA and mapping at a later time, it can be loaded via the layers control window. This option is particularly useful if you wish to analyze a number of clusters grouped together. Save them one at a time under different names, use your spreadsheet program to build them into one table, and reload this with the layers control menu.

4. Theory

The FloraMap system is based on calculating the probability that a climate record belongs to a multivariate normal distribution described by the climates at the collection points of a calibration set of organisms. It was designed for naturally occurring plant species; its use may be extended to cover the natural occurrence of any organism whose distribution is largely determined by climate. It uses a set of interpolated climate surfaces, a method for calculating the probability model, and a method for mapping the climate probabilities over the climate surface.

The Climate Surfaces

Spatially interpolated climate surfaces are now available for many areas. These usually handle long-term climate normals interpolated over a DEM by various methods (Hutchinson 1997, Jones 1991). Pixel size depends on the underlying elevation model. It may be as little as 90 m (Jones 1996), which results in a massive dataset, or 10 minutes of arc (about 18 km), which is as large as is practicable in many instances. In the latter case, the normal elevation model is the National Oceanographic and Atmospheric Administration (NOAA) TGPO006 (NOAA 1984). We have produced interpolated datasets at CIAT for Latin America and Africa using data from about 10 000 stations for Latin America and 7000 for Africa. Each set of surfaces consists of the monthly rainfall totals, monthly average temperatures, and monthly average diurnal temperature range. This makes 36 climate variates in three groups of 12.

We use a simple interpolation algorithm based on the inverse square of the distance between the station and the interpolated point. For each interpolated pixel we find the five nearest stations. Then the inverse distance weights are calculated and applied to each monthly value of the data type being interpolated. Thus, for five stations with data values $\boldsymbol{x}$ and distances from the pixel distance d:

$$x_{pixel} = \frac{1}{\sum_{i=1}^{5} d_i^{-2}} \times \sum_{i=1}^{5} \frac{x_i}{d_i^2} \qquad (1)$$

Temperature data are standardized to the elevation of the pixel in the DEM using a lapse rate model (Jones 1991).

Using this simple interpolation has various advantages. First, it is the fastest of all the common methods. Second, it puts the interpolated surface exactly through each station point, because the weight *1/(d(I)**2)* becomes infinite as *d* approaches zero. Third, the interpolation is highly stable in areas of sparse data. It approaches the mean of the nearest stations while they all become equally distant. Fourth, it is relatively stable against errors in station elevation; only the local region of that station is affected. On the other hand, laplacian spline techniques and co-Kriging both propagate these errors more extensively. This is one advantage of using a proven lapse rate model instead of fitting a local one, as do both of these latter techniques.

The method has two small disadvantages. First, the derivative of the surface becomes zero as it passes through the station point. In other words, each station is on a small plateau or step in the interpolated surface. This is usually much smaller than the pixel size and hence is not noticeable. Second, a (usually small) step occurs in the fitted surface as stations come into or drop out of the fitting window. Where the station density is high with respect to the pixel size, this is almost impossible to see. Where the stations are not so dense, it can produce unsightly straight lines or smooth arcs in the fitted rainfall data, which are not tied to elevation. Inspection of the surface's profile usually shows that these are negligible artifacts, but they are unsightly and can undermine confidence in the surface maps.

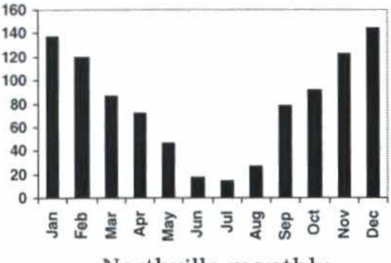
Climate Date Standardization (Rotation)

The climatic events that occur through the year, such as summer/winter and start/finish of the rainy season, are of prime importance when comparing one climate with another. Unfortunately, they occur at different dates in many climate types. The most obvious case is where climates are compared between points in the northern and southern hemispheres, but more subtle differences can be seen in climate event timing throughout the tropics. What we need is a method of eliminating these differences to allow us to make comparisons free of these annual timing effects.

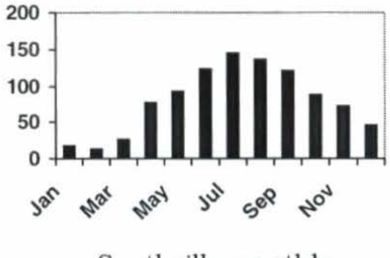
Let us look at two hypothetical climate stations. They are in a typical Mediterranean climate—warm wet winters, hot dry summers. Northville could be somewhere in California, and Southville might be in Chile. The August rainfall in Southville is received in January in Northville. If we plot these rainfalls in polar coordinates, we can readily see that to compare them we need to rotate them to a standard time.

Monthly rainfalls for Northville and Southville.

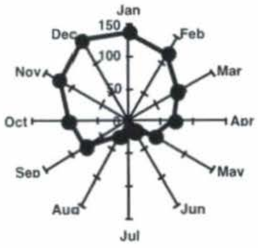
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Northville	137	120	87	72	46	18	14	27	78	92	123	145
Southville	18	14	27	78	92	123	145	137	120	87	72	46



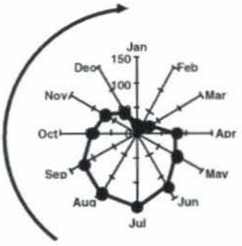
Northville monthly



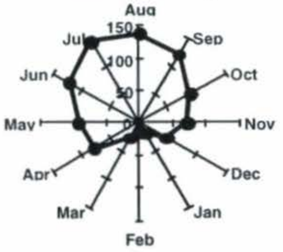
Southville monthly



Northville



Southville



Southville rainfall rotated to coincide with timing of Northville

How do we do this automatically? The answer is the 12-point Fourier transform. This is fortunately the simplest of all the possible Fourier transform algorithms. It is highly computationally efficient and fast. In fact, it is the basis of nearly all fast Fourier transform algorithms that break the problem down sequentially into the simple 12-point case. It takes the 12 monthly values and converts them to a series of sine and cosine functions. The one used in FloraMap has a modification to make it conserve the monthly total values (Jones 1987). The equation produced is:

$$r = a_0 + \sum_{i=1}^{6} a_i \sin(ix) + b_i \cos(ix) \qquad (2)$$

This can be rewritten as a series of frequency vectors, each with an amplitude α_i and a phase angle, θ_i:

$$\alpha_i = \sqrt{a_i^2 + b_i^2} \qquad \theta_i = \sin\left(\frac{b_i}{\alpha_i}\right) = \cos\left(\frac{a_i}{\alpha_i}\right) \qquad (3)$$

If we subtract the first phase angle from all the other vectors in the set then we have produced a rigid rotation of the vectors. This is the rotation that we are seeking. It puts the maximum of the first frequency at a phase angle of zero and places the rest in positions equivalent to their angular separation in the original data. We then use the first phase angle for rainfall to rotate the data for temperature and diurnal temperature range, and these variates are rigidly rotated along with the rainfall.

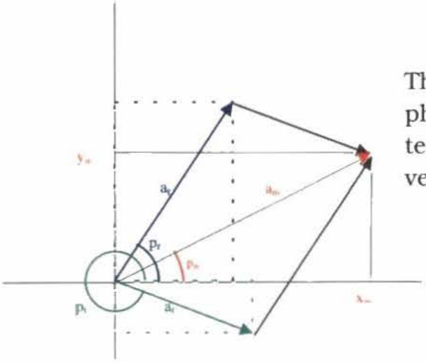
It is obvious how this algorithm works for climate records with unimodal rainfall. Climates could exist that are ambivalent with respect to a first frequency rotation. In practice, these are hardly ever met, the only serious case being where no rainfall occurs at all throughout the year.

This explanation works well for the tropics and almost everywhere as stated in the Release 1.0 FloraMap manual. There was a small chance of the procedure going off the rails if an accession set was fitted to a model in latitudes high enough to exhibit Mediterranean climates (as used in the example above). In the case when some of the accessions fall in the winter rainfall areas and some in strongly summer rainfall (non-Mediterranean) areas, the resulting model could have a very poor fit. Because this is botanically unlikely, it probably has not yet been observed in practice, although the case has arisen when running an artificial test set across the Andes in Chile/Argentina.

Dr Ian Makin of the International Water Management Institute (IWMI) kindly gave us access to the IWMI World Water and Climate Atlas to make climate grids to extend the range of FloraMap. We chose to try the grid for Europe because we have some potential users wanting to look at this area. The problem then arose. Temperature is by far the dominant climate determinate in Western Europe. The rainfall patterns can be winter, summer, or indeterminate over relatively short distances.

We therefore have the possibility of rotating on rainfall or temperature, but when to decide which is the dominant? We tried many combinations of rules, but unfortunately came to the conclusion that none were acceptable. They all resulted in a hard line across the map at some point where the rotation basis changed. This led to climates that should have been grading imperceptibly from one type to another suddenly jumping at a discontinuity, and would have given the users serious problems when fitting models in these areas.

The best solution found is to use BOTH the rainfall and the temperature in calculating the rotation phase angle. Thus:



The vector diagram of the first phases of rainfall (a_r) and temperature (a_t) with the resultant vector (a_m)

The resultant phase angle and amplitude are then:

$$y_m = a_r \cos p_r + a_t \cos p_t$$

$$x_m = a_r \sin p_r + a_t \sin p_t$$

$$a_m = \sqrt{y_m^2 + x_m^2}$$

$$p_m = \text{angle}\left({}^{x_m}\!/_{a_m}, {}^{y_m}\!/_{a_m} \right)$$

Unfortunately, this does not completely solve the problem of fitting a model to climates with different weather determinants. However, the vast majority of climates in the world are either:

(1) Rainfall determined where temperature is not an important seasonal effect (large areas of the tropics and subtropics),

(2) Temperature determined where rainfall is even throughout the year (most of the rest of the tropics and some temperate climates), or

(3) Rainfall and temperature determined when the two variates are highly correlated (summer rains - most of the rest of the world).

The Odd Man Out is:

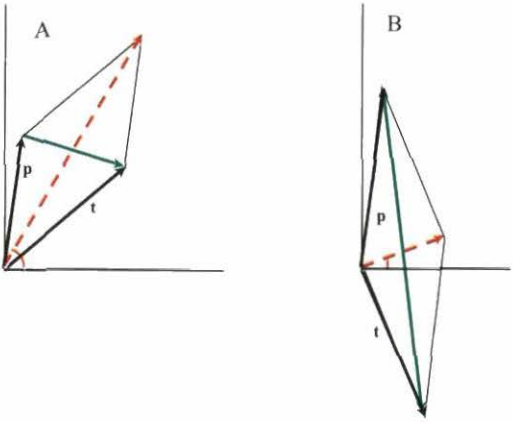
(4) Winter rains and hot dry summers (almost only Mediterranean climates).

Luckily, the Mediterranean climates are at moderately high latitudes and we can afford to have the rotation dominated by temperature without losing generality in the rotations and comparisons. We therefore need to increase the weighting for the temperature vector smoothly as we approach the Mediterranean climates (in order to avoid a sudden swing).

The following weightings were found to work well:

p = rainfall mm

t = temperature x 2 x abs(latitude)



There is a potential trap when the two vectors almost cancel each other. This could result in wild swings of the rotation angle for small changes in the rainfall and temperature vectors. This becomes more likely as the situation passes from that in A (above) to B and beyond. The dashed arrows are the rotation vectors as before, but calculated on the weighted rainfall and temperature vectors.

Where the rotation vector is the vector sum **r** + **t**, the counter-diagonal vector is the difference **r** – **t**. It can be readily seen that the dangerous areas will be when **r** – **t** is much greater than **r** + **t**. We can therefore use a handy index of stability, *s*.

$$s = \arctan\left(\frac{|\mathbf{r}-\mathbf{t}|}{|\mathbf{r}+\mathbf{t}|}\right)$$

This will be zero for stable states where the rotation angle is dominated by rainfall, by temperature, or by both acting in concert. It will approach $\pi/2$ as the vectors tend towards cancelling their effects. Because we can map this index we can check for areas where this indeterminate rotation might occur. Areas of relatively high *s* (potential instability) occur on the USA Pacific Coast, in Chile, northeastern Brazil, Sri Lanka, and through some areas of Central Africa. However, in no area does the index reach 80 degrees. Although this appears high, the phase angles are rotated correctly and in fact there is little chance of a spurious rotation.

If you are uncertain of the model fits when including accessions from these areas, please use the ClimateDiagram tool to investigate the situation. In the case of high precision grids, there may be the occasional pixel that rotates in an odd way and we will review this possibility when we create the new grids. However, for the present FloraMap grids there will be no problem.

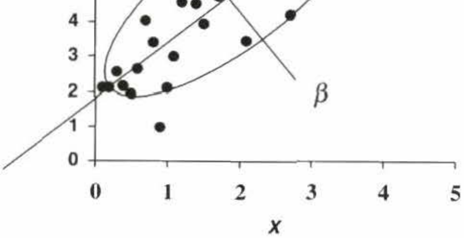
To save computing time, the whole climate surface is rotated according to these rules and all operations in FloraMap are done in the rotated phase space.

☞ *The only exception to this is when the user requests a climate diagram for an accession point or a climate surface point.*

The Model Calculations

Once we have rotated each record in the calibration set we are ready to construct the probability model. The data we are dealing with has 36 climate variates. If we used all of these in the form in which they are presented we would have the problem of constructing the probability model in 36-dimensioned space. Although not too difficult for a modern computer, this does present problems for the user when trying to visualize what is happening. The other consideration is that all of the climate variables can be highly correlated, making the probability model even harder to understand. A way around this problem is to use a principal components analysis (PCA). A PCA constructs sets of linear combinations of variates so as to maximize the variance in each from the original data. These linear combinations have another highly useful aspect. They are orthogonal to each other, completely uncorrelated, and so can be handled separately or in sets without unexpected interactions.

The operation can be illustrated in two dimensions as follows. The figure below shows a scatterplot of two variates x and y, quite highly correlated and therefore not at all independent. For any change in x we would expect a change in y. However, we can find two new axes, α and β, such that they are not correlated, and that the variance accounted for in the first of the new axes is maximized. Note that α is not the regression line of y on x and hence goes straight through the group of points.



In this case, a = $0.454x + 0.891y$ and b = $0.891x - 0.454y$. These new axes are orthogonal and uncorrelated. Movement along the a axis does not imply any movement at all along the b axis. The component a accounts for 95.6% of the original variance, b merely 4.4%. The trick to this linear transform is to calculate the eigenvalues and eigenvectors of the variance-covariance matrix of the system of variates. In FloraMap's case, this is a 36 x 36 matrix of climate variates.

In matrix notation we need to find a matrix **Q** and a diagonal matrix Λ such that:

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathrm{diag}\,\lambda = \Lambda \qquad (4)$$

where **A** is our variance-covariance matrix.

The matrix Λ, composed of the elements λ, is the diagonal matrix of the eigenvalues, which in our case hold the variance of the eigenvectors. The matrix **Q** is a symmetric matrix, which holds the eigenvectors as both rows and columns. The eigenvectors have two highly useful properties, one of which has been mentioned above—they are linearly independent of each other. The second useful property is that an eigenvector multiplied by any scalar is still an eigenvector.

The variance-covariance matrix does not have to be full rank for this operation. FloraMap will fit to as few as three calibration points. However, with accession sets of less than 15 points the results may not be reliable.

PCA can be performed on the sums of squares and cross products (SSCP) matrix, the variance-covariance matrix, or on the correlation matrix of a group of variates. In FloraMap, we use the variance-covariance matrix by standardizing the variates before we calculate the SSCP. But, we differ from many standard analyses in that our data has a structure that we want to preserve rather than standardize completely. The data are actually three groups of 12 values for different climate variables—rainfall, temperature, and diurnal temperature range. We want to conserve this difference to allow the user to apply weight across the board for the climate variables, for example, increasing the importance of rainfall over that of temperature. In addition, the information across the 12 monthly values is of critical interest and we do not wish to standardize it away. We therefore standardize all rainfall values by the common variance for rainfall and so forth.

☞ *At the time of writing, we standardize each monthly variate to zero mean. We are looking into the possibility of giving the option to standardize using only the group mean as well. This will give a more critical fit to effects throughout the year, but at present, unwanted effects translate through into the component values when the weights are changed.*

Once we have found the Λ and **Q**, we can describe the system of climate variates in terms of the principal components and their variances (eigenvectors and eigenvalues). We can choose a subset of the components (because the eigenvectors are independent), and we can scale them individually (because multiplying or dividing by a constant does not change the eigenvector's properties). This last point is important because this is exactly what we want to do to calculate the probabilities.

Probability Calculations

The normal probability density function for a single variate is given by:

$$z = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \tag{5}$$

From the integral of this function we can estimate the probability of observing a point drawn from this population. Traditionally, we look at the probability that a point might lie further from the origin than the point in question. We also usually estimate the distribution parameters from the sample that we are investigating. Because of this, we use other statistics such as **student's t** to estimate the probability.

☞ *In FloraMap, we make a simplifying assumption that the accessions calibration set will contain sufficient points so that estimating from the sample will be equivalent to knowing the population parameters. This will not be true for small calibration sets, and so the probabilities calculated will not be strictly accurate. However, provided the user recognizes this, the probabilities can still be used as a mapping index.*

For multiple dimensions with n independent (orthogonal) variates, the probability density function becomes:

$$z = \frac{1}{\sqrt{2\pi}} e^{-\left(t_1^2 + t_2^2 + \cdots t_n^2\right)/2} \tag{6}$$

The integral of this can be obtained by repeated integration, but specifying the integration bounds for each subsequent integration in terms of the previous functions is untidy and tedious. Here is an easier way to look at it. We want the probability that any point in a distribution falls within a radius of:

$$r = \sqrt{\left(t_1^2 + t_2^2 + \cdots t_n^2\right)} \tag{7}$$

The volume of a sphere of dimension n is:

$$\frac{r^n \sqrt{\pi^n}}{\Gamma\left(\frac{n+2}{2}\right)} \tag{8}$$

Note that as n increases, the volume of the sphere tends to zero. Thus the probability integral constructed in space with large n will be counter-intuitively small.

The volume of an infinitely thin shell of this sphere at radius r is:

$$\frac{n r^{n-1} \sqrt{\pi^n}}{\Gamma\left(\frac{n+2}{2}\right)} \tag{9}$$

The derivative of the probability integral at this shell is:

$$z = \frac{nr^{n-1}\sqrt{\pi^{n}}}{\Gamma\left(\frac{n+2}{2}\right)}e^{-r^{2}/2} \tag{10}$$

Therefore the integral from 0 to *r* is:

$$\frac{n\sqrt{\pi^{n}}}{\Gamma\left(\frac{n+2}{2}\right)}\cdot\int_{0}^{r} r^{n-1}e^{-r^{2}/2}\,dr \tag{11}$$

Taking only the portion to the right of the integral sign, and dividing by the limit as *r* passes to infinity from the left, we have, for even dimensions:

$$\frac{\lim}{r \to \infty}\int_{0}^{r} r^{n-1}e^{-r^{2}/2} = \Gamma\left(\frac{n}{2}\right)2^{(n-2)/2} \tag{12}$$

$$p = 1 - \frac{e^{-r^{2}/2}\left(r^{n-2} + (n-2)r^{n-4} + (n-2)(n-4)r^{n-6}\cdots\Gamma\left(\frac{n}{2}\right)2^{(n-2)/2}\right)}{\Gamma\left(\frac{n}{2}\right)2^{(n-2)/2}} \tag{13}$$

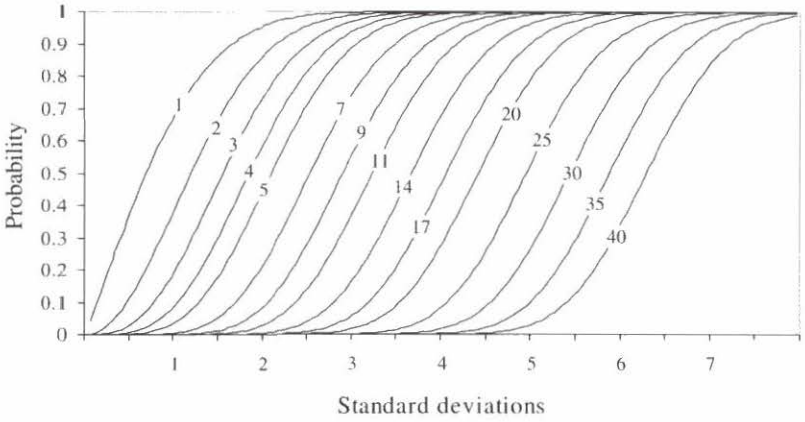
Factorizing, this becomes:

$$p = 1 - e^{-r^{2}/2}\left(\frac{\frac{\frac{\frac{\frac{1}{(n-2)}r^{2}+1}{(n-4)}r^{2}+1}{(n-6)}r^{2}+1}{\vdots}r^{2}+1}{(n-(n-2))}r^{2}+1\right) \tag{14}$$

And for odd dimensions, factorizing as we go:

$$\frac{Lim}{r\to\infty}=\frac{\sqrt{2}\cdot\sqrt{\pi}\cdot(3)(5)(7)\cdots(n-2)}{2} \tag{15}$$

$$p=\operatorname{erf}\left(\frac{\sqrt{2}r}{2}\right)-\frac{\sqrt{2}r\,\mathrm{e}^{1r^2/2}}{\sqrt{\pi}}\cdot\left(\frac{\frac{\frac{\frac{\frac{1}{(n-2)}r^2+1}{(n-4)}r^2+1}{(n-6)}r^2+1}{\vdots}r^2+1}{(n-(n-3))}r^2+1\right) \tag{16}$$

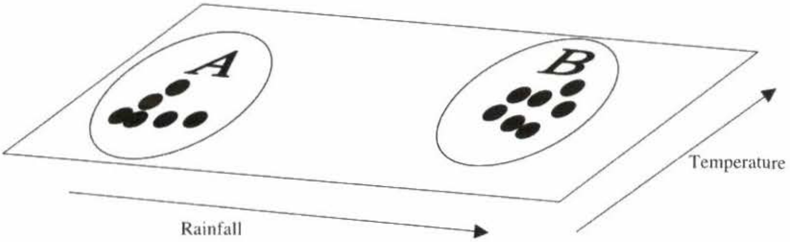


Probability integral in multiple dimensions: The probability of finding a point between the origin and radius *r* for N(0,1) populations in selected dimensions from 1 to 40.

This is an important result. If we did not have this, we could not maintain the correct level of probability as we passed from one set of dimensions to another. This is effectively what we do when we choose different sets of principal components.

Divergent Probabilities

Imagine a plane with rainfall varying from left to right and temperature varying from front to back.

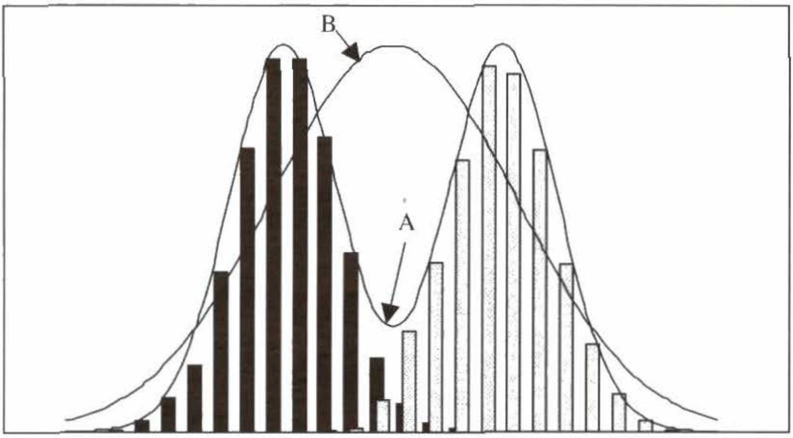


The points at (A) experience a dry cool regime, while the points at (B) experience a wetter and warmer climate. If we suppose that all are accession points of the same species, we would see no reason not to try to fit a single probability model to them. If it happens that the gap between them is purely fortuitous and exists merely because nobody has collected there, we might be right in fitting a simple, single model. On the other hand, if collectors have looked in the gap and found no accessions, then it would be wrong to fit a single model. Individual collectors may know which case pertains but, unfortunately, negative results rarely pass to germplasm collections. There is no entry for "I went there and did not find X", even though it is perhaps noted in the collector's field books.

FloraMap can give no clear-cut answers to a problem like this, but it can indicate that such a problem may be occurring, and we have included tools for the user to investigate such a possibility.

A useful indicator is the probability map itself. If the high probability areas consistently fall in areas with few accession points, while many accession points fall in medium to low probability areas, then a problem of the above type may exist. This seemingly unlikely result is, remarkably, not at all uncommon.

The figure below shows how it can occur. The two normal distributions are shown in the black and gray histograms. They are offset one from the other with just a little overlap in the middle. The bimodal line (A) is the sum of their normal distribution curves. It fits well to the histogram data and shows clearly the dip in the middle of the distribution where there are few observations. If we were to fit the total of the observations as if they were one continuous population, we would obtain the distribution curve (B). Note that the peak of the probability density function in this case occurs where least observations occur. It is a clear case of the meaningless mean. Try telling someone with one hand in scalding water and the other in ice that the average temperature is fine. Once he has treated his burns and frostbite he will be unlikely to thank you for your observation.



Two normal populations as histograms, showing the sum of their two distribution curves (A), and the distribution curve fitted to the full set of points (B).

The reason for this type of problem may be of a single type or, more often, of a combination of many reasons.

1. The species exists in the indicated areas, but has never been collected there.
2. Geographic or ecological barriers have prevented the spread of the species to these regions.
3. What was taken as a homogeneous group of germplasm at the start of the analysis is actually showing diverse groups of adaptation to climate, i.e., ecotypic differentiation.
4. There has been inadequate dispersal of recently emergent species.
5. There has been human interference.

In many situations, the elucidation of the reasons will have an important impact on the study and utilization of the germplasm.

A first check on the possibility of discontinuous distribution in climate space is the scattergram provided on the PCA window (see Chapter 3). This can view the accessions in any of the planes defined by the principal components. Because the components are orthogonal, with each slice of the component space you are getting a two-dimensional picture at right angles to all the other components. (Do not try to visualize where all the other right angles go to, it may give you a headache.)

The variance accounted for by each component falls off fast as you go down the list. Thus the slice to look at first is component 1 versus component 2. This often accounts for over three-quarters of all the variance. The first component is often considered as a 'size' component. The 'size' of a climate is a slightly difficult concept to handle, but by that we really mean a main trend component. That is to say, climates in one direction may be generally wetter and hotter, while those in the other direction are dryer and cooler. As you move down the components, they describe progressively more complicated and esoteric combinations of the data. However, it is sometimes possible to visualize the overall structure of a component down to the 4th or 5th and to give it a descriptive meaning. Sometimes the concepts of 'shape' or 'ratio' of variates come to mind. After the 5th, they can rarely be interpreted, but then the following components account for so little of the variance that they are usually disregarded.

Look at each scattergram slice for any clustering or discontinuities in the data. If obvious groupings occur, then this may support the evidence from a map showing the type of probability distribution described above. At this point it is wise to look back at the data and determine if any known genetic or morphological groupings occur within the accession sample that might explain the behavior of the model. FloraMap has a tool to assist with this analysis from the point of view of the climate data.

Cluster Analysis

We have incorporated various cluster methods into FloraMap to help develop a climate probability model that can cope with multiple populations. For further reading on cluster analysis we recommend Jain and Dubes (1988) for a much fuller treatment of the following descriptions. Everitt (1974) is an older treatment, but still very readable. Hartigan (1975) gives both a good practical description of the techniques and the Fortran source code of some interesting applications.

The methods we have incorporated are a small, but widely used, subset of clustering methods. They are single-link, complete-link, group average, weighted group average, unweighted centroid, weighted centroid, and Ward's method. Jain and Dubes class all of these as **SAHN** (sequential, agglomerative, hierarchic, and nonoverlapping). They are sequential because the elements are operated on one at a time, as opposed to all together. They are all agglomerative in that the clusters are built up stage by stage by adding members or by merging clusters. They are hierarchic, and a dendrogram tree can be constructed in all of them that shows the relationship between clusters at each level of clustering and between levels. The resultant clusters do not overlap in the *n* space in which they are drawn (in our case 36 dimensions). The distance measure is always an euclidean distance calculated from the climate data after transformation and weighting, but before the PCA. An euclidean distance is calculated as the square root of a sum of squares. In one case, Ward's method, this is a squared euclidean distance; we do not take the square root.

A further set of acronyms can be applied to some of the methods (see Jain and Dubes). The core of the acronyms, **PGM**, stands for **paired group methods**, the prefixes **U** and **W** for **unweighted** and **weighted**, and the suffixes **A** and **C** for **arithmetic mean** and **centroid**. Thus the group average method is often known as UPGMA, the weighted group average as WPGMA, the unweighted centroid as UPGMC, and the weighted centroid as WPGMC. These nomenclatures are sufficiently widely used to warrant inclusion in the method names in the cluster analysis window.

Lance and Williams (1967) suggested that a generalized formula for the distance matrix updating could cover the most common SAHN clustering method. Following Jain and Dubes for clusters *k*, *r*, and *s* this is:

$$d[k, (r, s)] = \alpha_r d[k, r] + \alpha_s d[k, s] + \beta d[r, s] + \gamma \, | d[k, r] - d[k, s] | \qquad (17)$$

Where d[] is the distance function, and d[*k*, (*r*, *s*)] is the distance from the newly formed cluster (*r*, *s*) and the existing cluster *k*, which has n_k members.

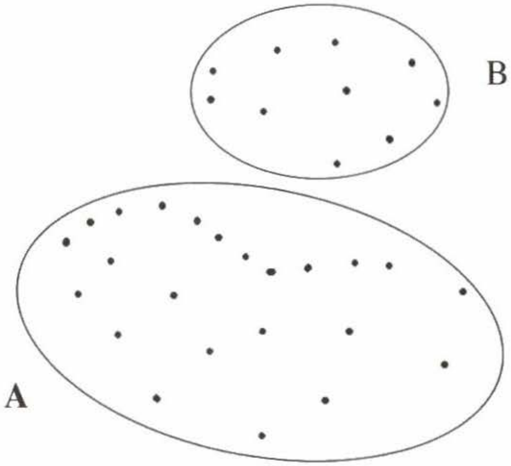
The following table shows the coefficients of the distance measures as used in the implementation in FloraMap, where n_r is the number of points in cluster *r*, n_s, in cluster *s*, and n_k in *k*.

Coefficient values for sequential, agglomerative, hierarchic, and nonoverlapping (SAHN) matrix updating algorithms (after Jain and Dubes 1988).

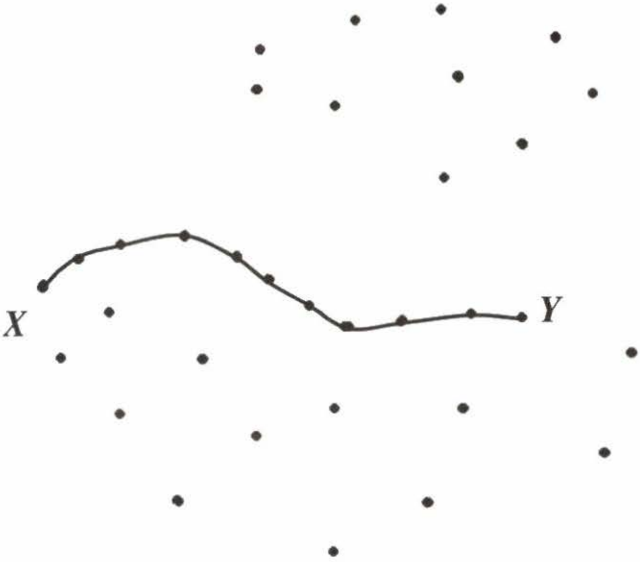
Clustering method	α_r	α_s	β	γ
Single-link	1/2	1/2	0	-1/2
Complete-link	1/2	1/2	0	1/2
UPGMA (group average)	$\frac{n_r}{n_r + n_s}$	$\frac{n_s}{n_r + n_s}$	0	0
WPGMA (weighted average)	1/2	1/2	0	0
UPGMC (unweighted centroid)	$\frac{n_r}{n_r + n_s}$	$\frac{n_s}{n_r + n_s}$	$\frac{-n_r n_s}{(n_r + n_s)^2}$	0
WPGMC (weighted centroid)	1/2	1/2	-1/4	0
Ward's method (minimum variance)	$\frac{n_r + n_k}{n_r + n_s + n_k}$	$\frac{n_s + n_k}{n_r + n_s + n_k}$	$\frac{-n_k}{n_r + n_s + n_k}$	0

The single-link method is closely related to the **minimum spanning tree** of graph theory. The same dendrogram may be generated from an agglomerative single-link algorithm ór by progressively removing the largest link from the spanning tree. The clusters therefore depend entirely on the distance between individual points and not on the properties of the emergent clusters. In practice, the method is fast and useful for pulling out many small clusters from a population where highly local clustering is expected. It has one characteristic that can be viewed as a strength, or a weakness, depending on the application.

Given the set of points below we may wish to draw attention to the two obvious clusters, (A) and (B).

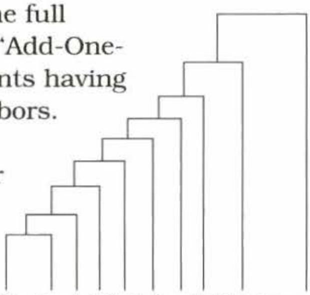


However, another clear grouping occurs that may have a physical meaning.



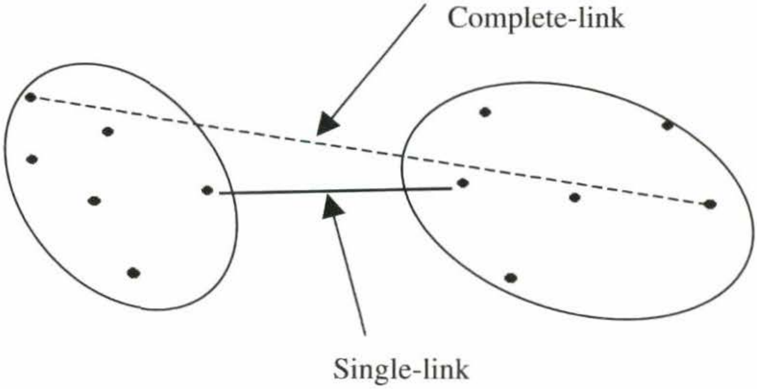
The group of points joined by the line *X* to *Y* might well be a transect amid many random points. Occasions could occur when the ability to separate out these data might be important. The single-link algorithm is ideal for this type of cluster, although it often fails on the more common ones. It tends to build clusters by adding on individual members, and produces dendrograms whose characteristic shape show this. It can be frustrating *if you* are looking for distinct clusters.

In some cases, the "Add-One-On" is a very useful tool. If we look at the last members incorporated in the full dendrogram they are almost always in the "Add-One-On" form. This means that they are the points having the greatest distance to their nearest neighbors. These are what we would commonly call outliers. It is a good bet that they are either points with location errors in their passports, or they are interesting accessions from rare environments (see TUTORIAL section, p 31).


The dreaded "Add-One-On" dendrogram.

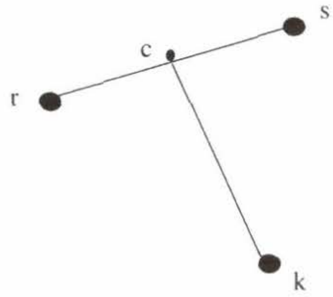
Most of the other methods that we have implemented are attempts to produce more compact clusters and avoid the dreaded "Add-One-On" dendrogram.

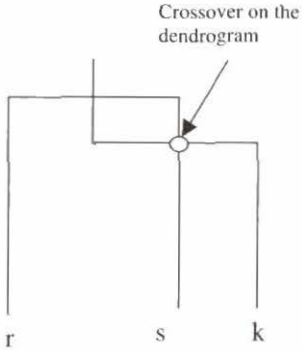
The *complete*-link method avoids this problem by measuring to the furthest members of two clusters to be joined.



The average methods are attempts to achieve the best characteristics of the two above, and use the arithmetic average of the two lines (dashed and solid) in the above diagram. UPGMA and WPGMA differ in that the unweighted option takes account of all the individuals in a cluster; the weighted method weights each cluster the same regardless of the number of individuals. Both UPGMA and WPGMA perform reasonably well with compact clusters, but can in some circumstances pick out clusters with odd shapes. This gives them an advantage over UPGMC and WPGMC, but the user needs to view the data carefully to ascertain if the clusters are real. A geometric representation cannot be given to the average methods, as there is no locality in the averaged distance.

UPGMC and WPGMC do, however, have a graphic representation. This can be used to illustrate an unfortunate consequence of the methods, as can be seen from the diagram below.





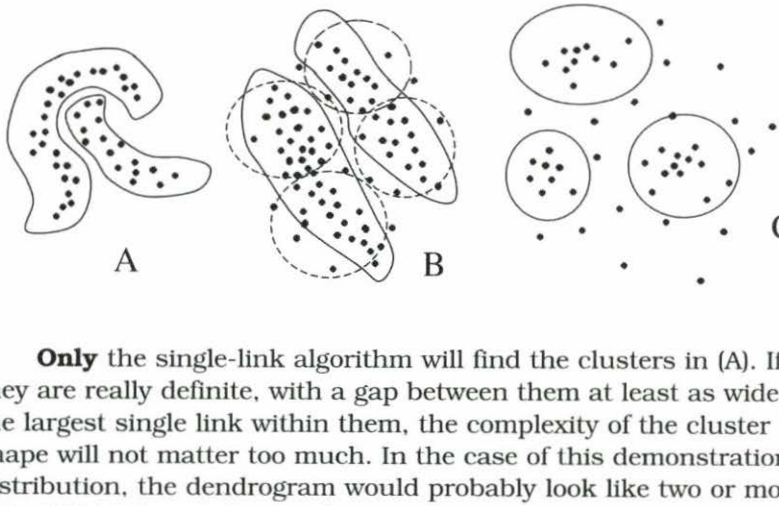
Here the clusters *r* and *s* have been joined to form a new cluster (*r, s*) with its centroid at *c*. The lines *r, k*, and *s, k* are longer than the line *r, s* and so the cluster (*r, s*) is formed prior to consideration of *k*. However, when cluster *k* is joined to the centroid of (*r, s*) the line *c, k* is shorter than the line (*r, s*). This means that the clusters are not formed in a monotonic scale and the dendrogram reflects this in having crossovers where it occurs as in the diagram to the right, above.

Although the dendrograms produced by UPGMC and WPGMC algorithms are difficult to interpret because of this effect, the methods usually produce compact, well-formed clusters. If the points are completely random, about 13% of cluster joins produce crossovers.

Ward's method produces the most compact clusters and the cleanest-looking dendrogram. This procedure attempts to minimize the variance within clusters and maximize that between cluster distances.

Care must be taken when interpreting any of the above methods. The compact clustering methods can produce compact clusters from completely random data, whereas the single-link algorithm is more likely to produce an "Add-One-On" dendrogram. Every chance should be taken to study the distribution of the data points. The PCA scattergram is one opportunity for doing this.

The following diagram shows some possible configurations and indicates which cluster techniques might produce the best results.



Only the single-link algorithm will find the clusters in (A). If they are really definite, with a gap between them at least as wide as the largest single link within them, the complexity of the cluster shape will not matter too much. In the case of this demonstration distribution, the dendrogram would probably look like two or more linked "Add-One-On" cascades.

To the naked eye, there appear to be two elongated elliptical clusters in example (B) (see the hand-drawn solid curves). Compact clustering algorithms like Ward's and the centroid methods will probably define the four clusters denoted by dashed ellipses.

The single-link and the Group average methods are more likely to find the elongated clusters. If the data were rescaled to tighten up the clusters, then the compact clustering methods would work well.

Dense spherical clusters in a background of random points, example (C), are best found by Ward's method. Great care must be taken to insure that the clusters produced are real and not an artifact of a random point distribution.

Bibliography

Carnahan B; Luther HA; Wilkes JO. 1969. Applied numerical methods. John Wiley, NY. 604 p.

Cooley WW; Lohnes PR. 1971. Multivariate data analysis. John Wiley, NY. 364 p.

Daly C; Taylor G. 1998. Annual maximum, minimum, and mean temperatures of the coterminous United States, 1961-1990. Water and Climate Center of the Natural Resources Conservation Service, April 1998. 12 GIS grids, GRASS format. Portland, OR.

Daly C; Taylor G. 1998. United States average monthly precipitation, 1961-1990. Water and Climate Center of the Natural Resources Conservation Service, April 1998. 12 GIS grids, GRASS format. Portland, OR.

Digby P; Galwey N; Lane P. 1989. Genstat 5. A second course. Clarendon Press, UK. 233 p.

Everitt B. 1974. Cluster analysis. Reviews of current research, no. 11. Social Science Research Council. Heinemann, UK. 122 p.

Hartigan JA. 1975. Clustering algorithms. John Wiley, NY. 351 p.

Hutchinson MF. 1997. ANUSPLIN Version 3.2 Users guide. The Australian National University. Centre for Resource and Environmental Studies, Canberra. 39 p.

Jain AK; Dubes RC. 1988. Algorithms for clustering data. Prentice Hall, NJ. 320 p.

Jones PG. 1987. Current availability and deficiencies data relevant to agroecological studies in the geographical area covered in IARCS. In: Bunting AH, ed. Agricultural environments: characterization, classification and mapping. CAB International, UK. p 69-82.

Jones PG. 1991. The CIAT climate database version 3.41. Machine readable dataset. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Jones PG. 1996. Climate database for Haiti. Machine readable dataset. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Jones PG; Thornton PK. 1993. A rainfall generator for agricultural applications in the tropics. Agric. For. Meteorol. 63:1-19.

Jones PG; Thornton PK. 1997. Spatial and temporal variability of rainfall related to a third order Markov model. Agric. For. Meteorol. 86:127-138.

Jones PG; Galwey N; Beebe SE; Tohme J. 1997. The use of geographical information systems in biodiversity exploration and conservation. Biodivers. Conserv. 6:947-958.

Jones PG; Rebgetz R; Maas BL; Kerridge PC. 1996. Genetic diversity in *Stylosanthes* species: a GIS mapping approach. Paper and poster presented at the First International Symposium on Tropical Savannas, 24-29 March, Brasilia, Brazil. 14 p. Typescript copies available from CIAT, Cali, Colombia.

Lance GN; Williams WT. 1967. A general theory of classificatory sorting strategies: II. Clustering algorithms. Computer J. 10:271-277.

Morrison DF. 1967. Multivariate statistical methods. McGraw-Hill Series in Probability and Statistics. McGraw-Hill, NY. 338 p.

NOAA (National Oceanographic and Atmospheric Administration). 1984. TGPO006 D. Computer compatible tape. Boulder, CO.

Press WH; Flannery BP; Teukolsky SA; Vetterling WT. 1986. Numerical recipes: the art of scientific computing. Cambridge University Press, UK. 818 p.

Appendix A

FloraMap File Types

ACP.files are accession point files composed as space-delimited ASCII files with column headings. They are the ASCII equivalent of the DBF accession point files. FloraMap can read them directly.

DBF files are used for various types of file associated with a map coverage. They can contain accession points data, climate data, or information about polygons in the shapefiles.

MAP files are FloraMap files that contain the information to associate DBF files, accession points files, and shapefiles in a map coverage. A MAP file contains references to the several map layers combined together to represent a map. FloraMap will construct a set of all the files necessary for the map and name them following the name you give to the MAP file. It also contains the visual attributes for every layer such as colors, font sizes, et cetera.

SHP files are shapefiles that delineate point, line, or polygon data. They are also compatible with ArcView.

SHX files are specialized index files that give meaning to the shapefiles.

TXT files are produced when a report file is saved. These are ASCII space-delimited data files.

An **ESRI shapefile** consists of a main file, an index file, and a dBASE table. The main file is a direct access, variable-record-length file in which each record describes a shape with a list of its vertices. In the index file, each record contains the offset of the corresponding main file record from the beginning of the main file. The dBASE table contains feature attributes with one record per feature. The one-to-one relationship between geometry and attributes is based on record number. Attribute records in the dBASE file must be in the same order as records in the main file.

Example

Main file:	counties.shp
Index file:	counties.shx
dBASE table:	counties.dbf

The three files above describe a map layer.

Creating Accession Points Files

Accession points files can be constructed as **ACP** or **DBF** files. They must contain at least two columns with the heading latitude longitude. No other data columns are absolutely necessary.

```
latitude longitude
″
5.6 -76.3
″
4.3 -74.3
10.4 -73.0
```

This is a minimal ACP file with four accession points.

If you present FloraMap with a file like this, without a column for elevation, then the analysis will work using the elevation taken from the Climate Database grid, even if you have specified correct temperature in the configuration window. To allow for correcting the temperature by elevation you must provide a column to hold the elevation data. You can also add as many columns of other data as you wish. They will be read in with the accession points and kept as identification with the accession points throughout the analysis.

```
latitude longitude elevation  id   info
5.6       -76.3       1200     1   x
4.3       -74.3        600      2   v
10.4      -73.0        200      3   bbb
-20.3     -44.3        120      4   uuvv
```

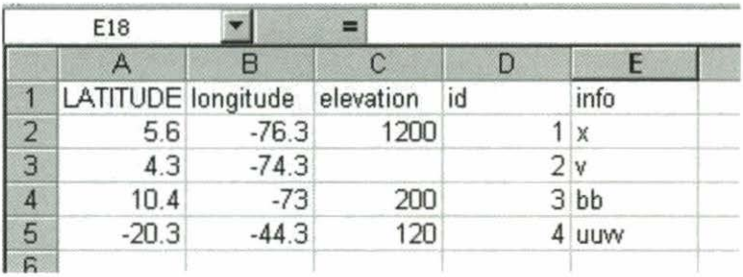
You can add as many columns as you like as long as they are complete.

If you have data columns that are not complete for all accessions then you cannot use this space-delimited format. You should prepare the file in a comma-delimited format allowing for the missing values. Thus:

```
LATITUDE, longitude, elevation, id, info
5.6, -76.3, 1200, 1, x
4.3, -74.3,     , 2, v
10.4, -73.0, 200, 3, bb
-20.3, -44.3, 120, 4, uuvv
```

The elevation for point 2 now has a missing value for elevation.

FloraMap cannot read this comma-delimited file. Read it in to your spreadsheet program, check that the formatting is correct, and save it as a dBASE 4 file, with file extension .dbf.



	A	B	C	D	E
1	LATITUDE	longitude	elevation	id	info
2	5.6	-76.3	1200	1	x
3	4.3	-74.3		2	v
4	10.4	-73	200	3	bb
5	-20.3	-44.3	120	4	uuvv
6					

Index