

## 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  $x$  and distances from the pixel distance  $d$ :

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

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)**

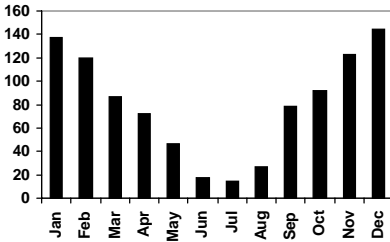
The climatic events that occur through the year, such as summer/winter and start/finish of the rainy season, are of prime importance when comparing one climate with another. Unfortunately, they occur at different dates in many climate types. The most obvious case is where climates are compared between 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.

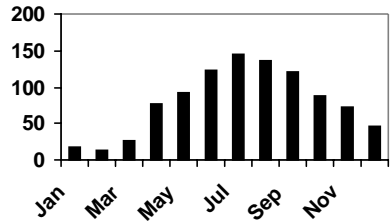
Let us look at two hypothetical climate stations. They are in a typical Mediterranean climate—warm wet winters, hot dry summers. Northville could be somewhere in California, and Southville might be in Chile. The August rainfall in Southville is received in January in Northville. If we plot these rainfalls 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.**

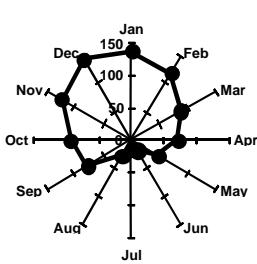
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Northville	137	120	87	72	46	18	14	27	78	92	123	145
Southville	18	14	27	78	92	123	145	137	120	87	72	46



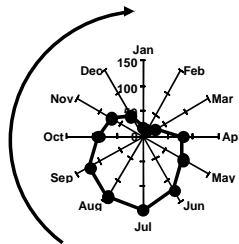
Northville monthly



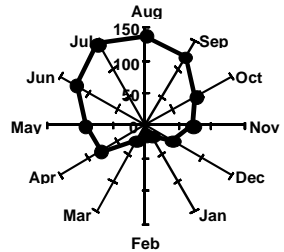
Southville monthly



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:

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

This can be rewritten as a series of frequency vectors, each with an amplitude  $a_i$  and a phase angle,  $\alpha_i$ :

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

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.

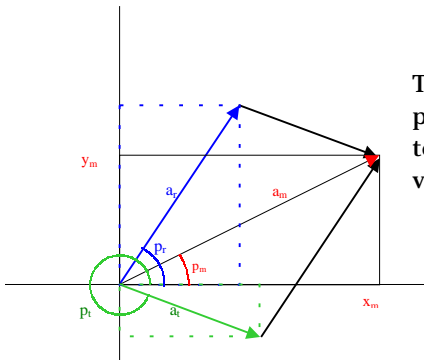
It is obvious how this algorithm works for climate records with unimodal rainfall. Climates could exist that are ambivalent with respect to a first frequency rotation. In practice, these are hardly ever met, the only serious case being where no rainfall occurs at all throughout the year.

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:

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

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

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

$$p_m = \text{angle} \left( \frac{x_m}{a_m}, \frac{y_m}{a_m} \right)$$

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:

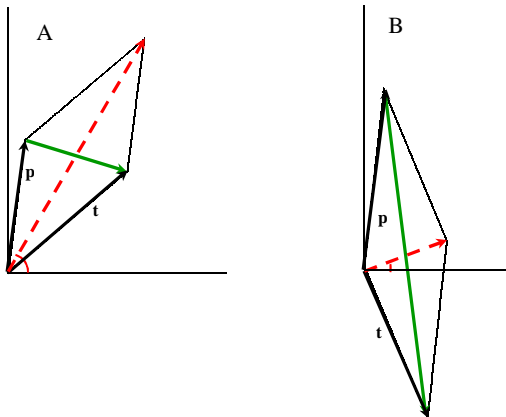
- (4) Winter rains and hot dry summers (almost only Mediterranean climates).

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  $\mathbf{r} + \mathbf{t}$ , the counter-diagonal vector is the difference  $\mathbf{r} - \mathbf{t}$ . It can be readily seen that the dangerous areas will be when  $\mathbf{r} - \mathbf{t}$  is much greater than  $\mathbf{r} + \mathbf{t}$ . 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.

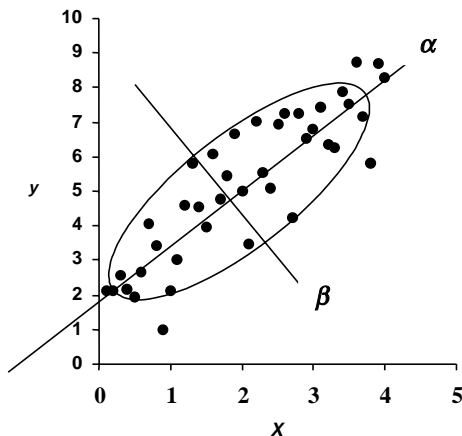


*The only exception to this is when the user requests a climate diagram for an accession point or a climate surface point.*

## 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-dimensional 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,  $a$  and  $b$ , such that they are not correlated, and that the variance accounted for in the first of the new axes is maximized. Note that  $a$  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  $\Lambda$  such that:

$$Q^{-1}AQ = \text{diag } \lambda = \Lambda \quad (4)$$

where  $A$  is our variance-covariance matrix.

The matrix  $\Lambda$ , composed of the elements  $\lambda$ , 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  $L$  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:

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

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:

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{(t_1^2 + t_2^2 + \dots + t_n^2)}{2}} \quad (6)$$

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:

$$r = \sqrt{(t_1^2 + t_2^2 + \dots + t_n^2)} \quad (7)$$

The volume of a sphere of dimension  $n$  is:

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

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:

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

The derivative of the probability integral at this shell is:

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

Therefore the integral from 0 to  $r$  is:

$$\frac{n\sqrt{\pi}}{\Gamma\left(\frac{n+2}{2}\right)} \int_0^r r^{n-1} e^{-r^2/2} dr \quad (11)$$

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:

$$\lim_{r \rightarrow \infty} \int_0^r r^{n-1} e^{-r^2/2} = \Gamma\left(\frac{n}{2}\right) 2^{(n-2)/2} \quad (12)$$

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

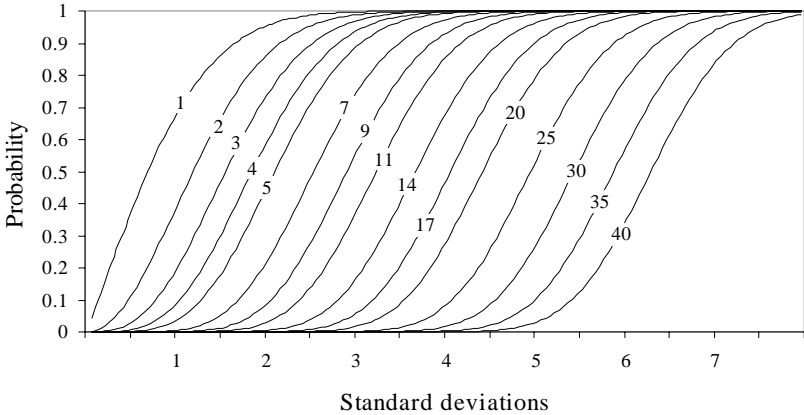
Factorizing, this becomes:

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

And for odd dimensions, factorizing as we go:

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

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

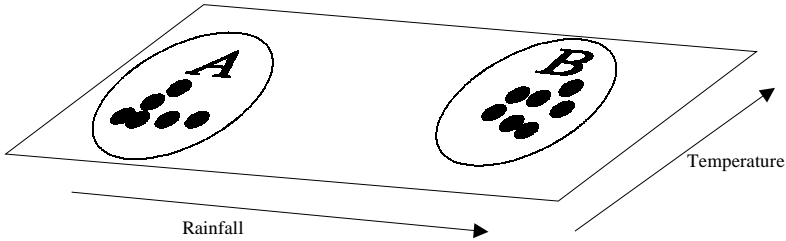


**Probability integral in multiple dimensions: The probability of finding a point between the origin and radius  $r$  for  $N(0,1)$  populations in selected dimensions from 1 to 40.**

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.



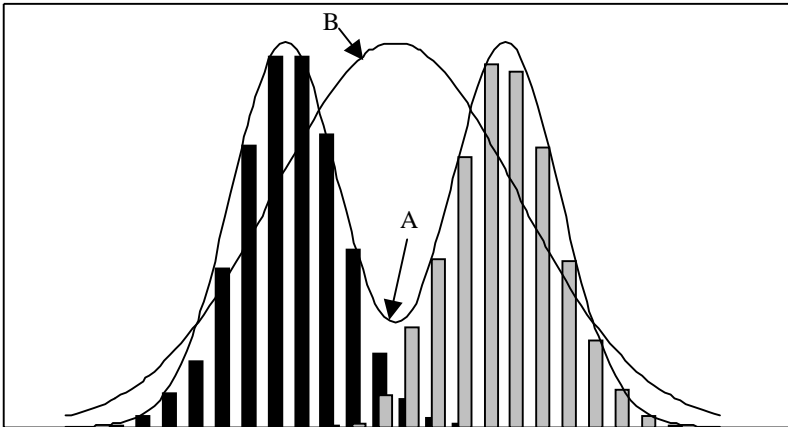
The points at (A) experience a dry cool regime, while the points at (B) experience a wetter and warmer climate. If we suppose that all are accession points of the same species, we would see no reason not to try to fit a single probability model to them. If it happens that the gap between them is purely 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 4<sup>th</sup> or 5<sup>th</sup> and to give it a descriptive meaning. Sometimes the concepts of 'shape' or 'ratio' of variates come to mind. After the 5<sup>th</sup>, 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:

$$d[k, (r, s)] = a_r d[k, r] + a_s d[k, s] + b d[r, s] + g |d[k, r] - d[k, s]| \quad (17)$$

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.

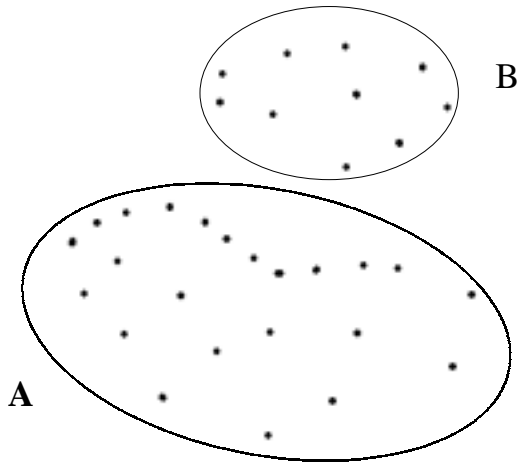
The following table shows the coefficients of the distance measures as used in the implementation in FloraMap, where  $n_r$  is the number of points in cluster  $r$ ,  $n_s$ , in cluster  $s$ , and  $n_k$  in  $k$ .

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

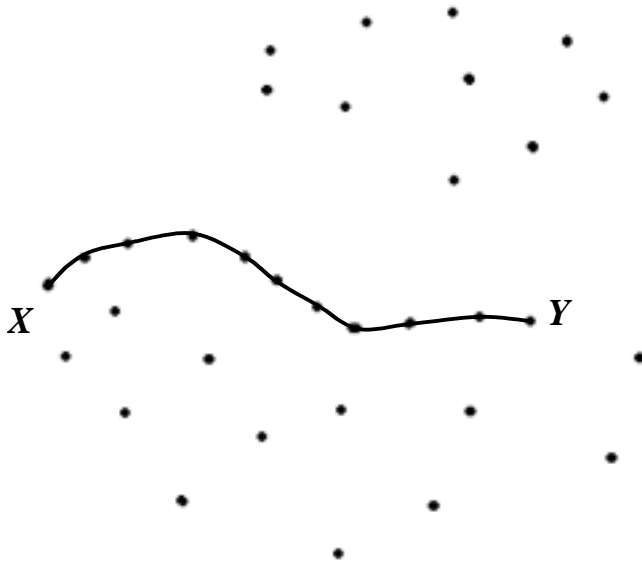
Clustering method	$a_r$	$a_s$	$b$	$g$
Single-link	1/2	1/2	0	-1/2
Complete-link	1/2	1/2	0	1/2
UPGMA (group average)	$\frac{n_r}{n_r + n_s}$	$\frac{n_s}{n_r + n_s}$	0	0
WPGMA (weighted average)	1/2	1/2	0	0
UPGMC (unweighted centroid)	$\frac{n_r}{n_r + n_s}$	$\frac{n_s}{n_r + n_s}$	$\frac{-n_r n_s}{(n_r + n_s)^2}$	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 or 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).

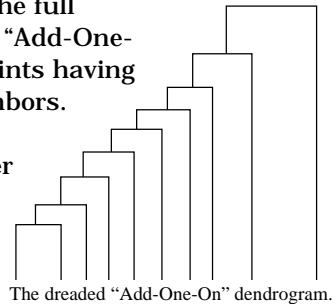


However, another clear grouping occurs that may have a physical meaning.



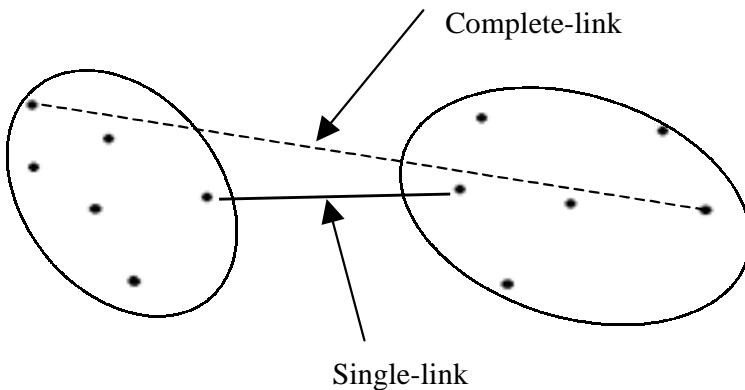
The group of points joined by the line X to Y might well be a transect amid many random points. Occasions could occur when the ability to separate out these data might be important. The single-link algorithm is ideal for this type of cluster, although it often fails on the more common ones. It tends 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).



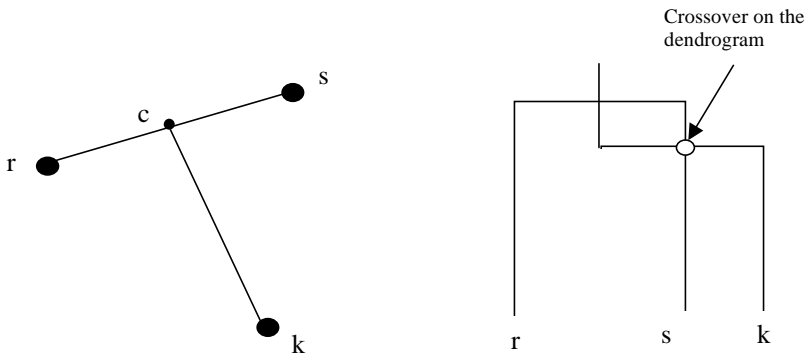
Most of the other methods that we have implemented are attempts to produce more compact clusters and avoid the dreaded “Add-One-On” dendrogram.

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.

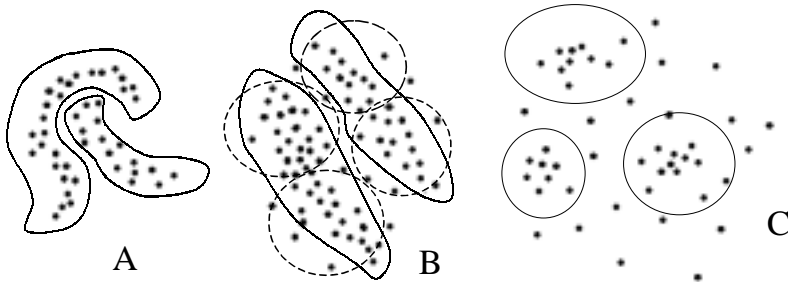
Although the dendrograms produced by UPGMC and WPGMC algorithms are difficult to interpret because of this effect, the methods usually produce compact, well-formed clusters. If the

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.