Cluster analysis applied in mapping the genetic divergence of populations of Araucaria angustifolia (Bert.) O. Ktze by isoenzymatic markers

RICARDO ALEXANDRE VALGAS
Programa de Pós-Graduação em Métodos Numéricos em Engenharia
Universidade Federal do Paraná
Centro Politécnico- Curitiba - Paraná
BRAZIL
ricardo.valgas@ufpr.br
ANSELMO CHAVES NETO
Programa de Pós-Graduação em Métodos Numéricos em Engenharia
Universidade Federal do Paraná
Centro Politécnico- Curitiba - Paraná
BRAZIL
OSMIR JOSE LAVORANTI
Empresa Brasileira de Pesquisa Agropecuária
EMBRAPA Florestas
BRAZIL
VALDERES APARECIDA DE SOUSA
Empresa Brasileira de Pesquisa Agropecuária
EMBRAPA Florestas
BRAZIL

Abstract: - This paper presents the methodology and results for the genetic mapping in five natural populations of Araucaria angustifolia (Bert.) O. Ktze trees by isoenzymatic markers. The areas of study in Brazil were Campos do Jordão – SP, Irati – PR, Caçador – SC, Camanducaia – MG and Barbacena – MG. For each tree six locus were evaluated: GOT-B, GOT-C, PGMA-A, MDH-B, SKDH-B, 6PGDH-B and alleles observed for endosperms and embryos (homozygote and heterozygote). The multivariate statistical technique used to provide the dendograms was Cluster Analysis using Euclidean and Manhattan distances, beyond Ward’s hierarchical method linkage, Single linkage and Average linkage to perform the dendograms. To validate the groupings were used MANOVA and $T^2$ Hotteling test. It was observed that the populations of Irati and Caçador are the most homogeneous group and Camanducaia, Barbacena and Campos do Jordão are the other group.

Key-Words: - Araucaria angustifolia, Isoenzymatic Markers, Genetic Mapping, Cluster Analysis, MANOVA, $T^2$ Hotteling Test.

1 Introduction

Many biological communities have been developed over thousands of years now are being devastated by human action. Therefore, many species are at extinction risk in this and in coming decades. The main cause of this problem is the growth of human population, causing negative environmental impacts by indiscriminate exploitation, conversion of forests into pastures and agricultural lands. In natural preserved habitats as natural reserves or national parks, special care is necessary to prevent the extinguishing of the remaining species, therefore the “in situ” conservation allows the continuity of the natural species evolution. In this context the conservation biology appeared, which investigates the negative impacts of the human being activities on ecosystems and develops practices that prevent the extinguishing of species, reintegrating them, when possible, to functional ecosystems. In Brazil, the natural forests have been the spot of the conservation biology due to its species richness and due to constant threats to its ecosystems. The necessity of conserving the remaining forests is urgent. For this procedure to be accomplished correctly, biological and ecological knowledge related to the populations to be preserved are necessary. Parameters such as the amount and distribution of the genetic variability, species reproductive biology, genic flow, regeneration, are,
among others, essential information in this process [1]. In Brazil we find a great reserve of natural resources, being considered the country of the biggest biodiversity of the planet due to climate and geomorphological variations found here, varying since Atlantic Forest, Brazilian savannahs, savannahs, caatingas, until the Amazonian Forest. Tropical forests cover approximately 20% of the terrestrial area (around 17 millions km²) and shelter more than a third of the planet species. During centuries, Brazilian tropical forests were considered inexhaustible sources of raw material for the most diverse purposes, such as: wood, fruits, fibers, among other products. However, these resources were exploited in a predatory way and most of them are found to be at extinction threatening.

2 Cluster analysis

The union of individuals in groups (clusters) based on the similarity or distance among them, considering the variables to be grouped, requires the application of a statistic technique of appropriated classifications designated to divide in subgroups a group of observed data [2]. To validate the clusters found means to guarantee that in fact the groups differ among each other. Such verification can be accomplished applying tests starting from univariate statistic tests for comparison of averages up to multivariate tests, such as, for example, the MANOVA (Multivariate Analysis of Variance) and the \( T^2 \) Hotteling test. Through MANOVA can be verified if there is significant statistic dissimilarities among the sample averages vectors of each cluster, that is, having as a null hypothesis \( (H_0) \), that the sample average vectors do not differ among each other: \( H_0: \mu_1 = \mu_2 = \ldots = \mu_k \). Table 1 summarizes some information.

<table>
<thead>
<tr>
<th>Sources of Variation</th>
<th>Matrix sum of squares</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>( B = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2 )</td>
<td>( k - 1 )</td>
</tr>
<tr>
<td>Error</td>
<td>( W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}<em>j)(x</em>{ij} - \bar{x}) )</td>
<td>( \sum_{j=1}^{k} n_j - k )</td>
</tr>
<tr>
<td>Total</td>
<td>( B + W )</td>
<td>( \sum_{j=1}^{k} n_j - 1 )</td>
</tr>
</tbody>
</table>

Table 1. MANOVA

The \( T^2 \) Hotteling test compares if the dissimilarity among the sample average vectors of two clusters formed by categorical independent variables is significant through the test statistic:

\[
T^2 = \left[ (\bar{x} - \bar{x}_-)(\bar{x} - \bar{x}_+)^T \right] \left( (\frac{1}{n_1} + \frac{1}{n_2}) S \right)^{-1} \left[ (\bar{x} - \bar{x}_-)(\bar{x} - \bar{x}_+)^T \right] \left( (\frac{1}{n_1} + \frac{1}{n_2}) S \right)^{-1} \left( (\bar{x} - \bar{x}_+)(\bar{x} - \bar{x}_-)^T \right) \tag{1}
\]

where \( S_p \) is the estimate of the conjoined covariance matrix. The \( T^2 \) statistics has an \( F \) distribution with \( p \) degrees of freedom in the numerator and \( n_1 + n_2 - p - l \) degrees of freedom in the denominator. In this case the null hypothesis \( (H_0) \) is evaluated using the test statistic known as Wilks’ Lambda (\( \Lambda \)), given by the expression:

\[
\Lambda = \frac{|W|}{|B + W|} \tag{2}
\]

where:

- \( B \): is the matrix that summarizes the variation among the effect of the treatments;
- \( W \): is the matrix that summarizes the variation in the clusters.

When applying the multivariate tests for the comparison of the averages is necessary that the random samples are independent and originated out of Gaussian populations, in other words, \( X_i \sim N_p(\mu_i, \Sigma) \), \( i = 1, \ldots, n \). The premise of Gaussianity for the vector \( X \) can be verified through a statistic test. In the case of the bivariated normal distribution with \( p = 2 \) variables and established \( \alpha \) significance level is obtained:

\[
p \left( \sum_{i=1}^{k} (X_j - \mu_j) \right) \leq \chi^2_{2} (1 - \alpha) = 1 - \alpha \tag{3}
\]

where \( \mu \) is replaced by the estimate \( \bar{X} \) and \( \Sigma \) by estimate \( S \).

2.1 Measures of dissimilarities

According to Johnson and Wichern [3], the individuals gathered in clusters represent a “proximity” measurement through some type of distance, in most cases.

2.1.1 Euclidean distance

It is certainly the best known and most used distance to calculate the distance between two objects in the \( p \)-dimensional space. Considering two vectors of the \( x \) and \( y \) matrices of observations, the Euclidian distance between them is defined by:
2.1.2 Manhattan distance
The Manhattan distance measures the distance that would be traveled to get from one data point to the other if a grid-like path is followed.

\[
d(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}
\]  

(4)

2.2 Hierarchical Clustering Methods
The first of these techniques is called Hierarchical Clustering Method. This one uses a matrix of distances to identify the pair of the most similar individuals that are gathered and considered a sole individual. From then, it is defined a new matrix of distances in which the most similar pair will be identified to form a new cluster, and then successively until that all the individuals are included in a same cluster. The Hierarchical Clustering Method is described in stages by Johnson and Wichern (1998) and consists of a systematic study, from the stage 1.

**Stage 1** The study starts with \( g \) clusters, being each one formed by a sole individual. The symmetric matrix of distances is calculated \( D = (d_{ij}) \) of dimension \( n \times n \), where \( d_{ij} \) represents the distance from the individual \( i \) to the individual \( j \):

\[
D = \begin{bmatrix}
d_{11} & d_{12} & \cdots & d_{1n} \\
d_{21} & d_{22} & \cdots & d_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
d_{n1} & d_{n2} & \cdots & d_{nn}
\end{bmatrix}
\]  

(6)

**Stage 2** In the symmetric matrix of distances \( D \) is the pair of the most similar clusters is found (short distance) represented by \( d_{AB} \) to indicate that \( A \) and \( B \) are the closest clusters.

**Stage 3** It appears, then, a new cluster denominated \( AB \). Now, the entries of the matrix of distances are updated, eliminating the lines and columns related to the \( A \) and \( B \) clusters, adding up a line and a column containing the distance from the \( AB \) group to the other remaining clusters.

**Stage 4** The stages 2 and 3 should be repeated \( g - 1 \) times (until all the individuals are included in a sole cluster) observing the clusters’ identities that are being formed and the level that they merge.

### 2.2.1 Ward’s Hierarchical Clustering Method
Johnson and Wichern [3] describe that this method is based on the “minimum information loss” when new clusters are formed, using the error sum of squares (ESS) criterion, between two grouping, for all the samples.

\[
ESS = \sum_{i=1}^{n} (x_i - \bar{x})'(x_i - \bar{x})
\]  

(7)

where \( \bar{x} \) represents the average of the samples and \( x_i \) a multivariate measure associated to the \( i \)th observation.

### 2.2.2 Single Linkage
In this method the distance between two clusters is represented by the shorter distance (or biggest similarity) among all the possible combinations of two pairs of individuals. Thereby, the first cluster is formed, for example, \((AB)\), and the distance between \((AB)\) to any other \( C \) individual is calculated by the expression:

\[
d_{(AB),C} = \min\{d_{AC}, d_{BC}\}
\]  

(8)

### 2.2.2 Average Linkage
In this procedure the considered distance is the average distance among all the possible pairs of individuals. Then, from the matrix of distances \( D = (d_{ij}) \) the elements with shorter distance are grouped, for example, \( A \) and \( B \) form a cluster if \( d_{AB} = \min\{d_{ij}\}, \forall i, j \) and the distance from \((AB)\) to \( C \) is expressed by:

\[
d_{(AB),C} = \frac{\sum_{i} \sum_{j} d_{ij}}{n_{(AB)} n_{C}}
\]  

(9)

where \( n_{(AB)} \) and \( n_{C} \) represent the quantities of elements present in each cluster.

### 3 Genetic Mapping
The trees used as sample for this study were obtained in the largest areas of environmental protection area located in the states of São Paulo (state park) and Minas Gerais (private área and PPA – Permanent Protection Area).
Protection Área), Paraná and Santa Catarina (in national parks), comprehending the cities of Campos do Jordão – SP, Irati – PR, Caçador – SC, Camanducaia – MG (PPA) and Barbacena – MG (private property).

3.1 Electrophoresis

The electrophoresis procedure was carried out starting from an amount of 20 seeds from each one of the Campos do Jordão trees and 8 seeds from the remaining population trees. The seeds were stored in a – 22°F until the enzyme extraction for the electrophoresis. Embryos and female gametophyte were homogenized separately in a tampon extraction containing 5g of Saccharose, 150mg of DTT (Dithiothreitol) and 3g of PVP (Polivinilpirrolidona) dissolved in 100ml of tampon 0,1M Tris-HCl, in a 7,5 pH. Enzyme extracts obtained from the female gametophyte and from the embryo were placed in gel in an adjacent way. The horizontal starch gel (10,5% of starch and 2,5 – 3,5% saccharose) was prepared according to Feret and Bergmann [4], Conkle et al. [5] and Liengsiri et al. [6]. The electrophoresis tampon system used was ASHTON adjusted to pH 8,6 for GOT and PGM, and the TRIS-Citrato system adjusted to pH 7,3 for the other enzymes. The gels have run at 80mA for the ASHTON system and at 180mA for the TRIS-Citrato system. Six locus were examined: GOT-B, GOT-C, PGM-A, SKDH-B, MDH-B and 6-PGDH-B. In total were prepared 2.280 zymograms. The details about the embryos electrophoresis methodology and gametophytes and the zymograms genetic interpretation are described in Sousa et al. [7].

3.2 Dendograms

Using the Euclidian and the Manhattan distance the respective matrices are obtained and after, the dendograms, keeping the three linkage methods. In Figure 1 the populations of Irati and Caçador have gathered forming the most homogeneous cluster, since they have the less distance (0,1152). Another cluster observed resulted from the gathering of the populations of Campos do Jordão and Camanducaia, with distance of 0,3755. Barbacena is very similar to this last cluster.

In Figure 2, where the Single linkage method is used, it was observed exactly the same grouping of the Figure 1, only with a little scale difference.

In Figure 3 again it was observed that the most homogeneous grouping formed by the populations of Irati and Caçador with distance of 0,2291. The other cluster observed now is formed by the populations of Camanducaia and Barbacena, being the population of Campos do Jordão very similar to this cluster.
To validate the grouping formed, a Hotteling $T^2$ test was accomplished comparing if there is a significant dissimilarity between the sample average vector of the cluster composed by the populations of Irati and Caçador with the sample average vector of the cluster formed by the populations of Campos do Jordão, Camanducaia and Barbacena. The matter of the Gaussianity of the random vectors observed was verified applying the test based on a chi-square statistics described previously. The p-value of the test accomplished with 95% of reliability is below 0.05 and the null hypothesis rejected, therefore there is a significant difference among the sample average vectors of the two studied clusters. Besides this comparison, it was also accomplished an analysis inside the clusters. Surely Irati and Caçador form a cluster with more similarities. For the group composed by the populations of Campos do Jordão, Camanducaia and Barbacena, at the level of 95% of reliability MANOVA provided a p-valor above 0.05, indicating that there is no significant difference among the sample average vectors of each of these populations.

4 Conclusion
This paper provides a piece of contribution to the knowledge of the genetic structure in populations of *Araucaria angustifolia* (Bert.) O. Ktze and creates technical-scientific information that make possible the lineation of strategies of preservation and rational profit for the sustained extractivist management. This paper demonstrates, also, that grouping analysis is a multivariate statistics technique able to produce good results in mapping genetic divergence in that kind of population, identifying those that have a genetic characteristic more similar through the cluster formation. Either MANOVA or $T^2$ Hotteling test may be used to evaluate the quality of the grouping formed. Besides, the necessary effort for araucarias preservation in those five regions considered may be concentrated to decrease costs and increase its efficiency.

References: