Apidologie, 1982, **13** (1), 15-20.

REPRÉSENTATION GRAPHIQUE DE POPULATIONS MULTINORMALES PAR DES ELLIPSES DE CONFIANCE

J.-M. CORNUET
Station de Zoologie, I.N.R.A., 84140 Montfavet

RÉSUMÉ

Dans le but essentiel de faciliter l'interprétation des analyses factorielles, nous présentons une méthode de représentation graphique de populations multinormales par des ellipses de confiance calculées à partir des moyennes et de la matrice de covariances de ces populations.

INTRODUCTION

En biométrie de l'Abeille, l'étude d'une nouvelle population passe le plus souvent par une étape de comparaison à des échantillons de référence représentant des races ou des écotypes connus. La procédure employée habituellement (CORNUET *et al.*, 1975; GADBIN *et al.*, 1979; CORNUET *et al.*, 1982) consiste à situer les colonies inconnues sur les premiers plans d'une analyse factorielle discriminante effectuée sur la base des populations de référence. Ces dernières forment des nuages plus ou moins distincts de points-colonies dont l'effectif total atteint souvent un nombre important (généralement entre 100 et 200). Il en résulte une représentation graphique assez confuse qui rend laborieuse l'interprétation des résultats. C'est pourquoi, dans le but d'alléger ces représentations, nous avons pensé remplacer l'ensemble des points-colonies de référence par des ellipses qui, au prix de l'hypothèse de multinormalité des populations, engloberaient statistiquement une proportion donnée des individus.

RAISONNEMENT

Considérons une population de n individus caractérisés par k variables. Nous associons à chaque individu r un vecteur aléatoire X_r de dimension k en faisant les hypothè-

ses classiques que ces n vecteurs aléatoires sont indépendants et suivent une même loi de probabilité d'espérance μ et de matrice de covariance Σ.

Soit X. le vecteur moyenne $(\mathrm{X}. = \frac{1}{n}\sum_r \mathrm{X}_r)$.

La matrice aléatoire $\mathrm{S} = \frac{1}{n-1}\sum_r (\mathrm{X}_r - \mathrm{X}.)^t (\mathrm{X}_r - \mathrm{X}.)$ est un estimateur non biaisé de Σ.

Si nous supposons que la loi de probabilité des X_r est une loi multinormale, nous savons alors que la variable $n^t(\mathrm{X}_r - \mu)\mathrm{S}^{-1}(\mathrm{X}_r - \mu)$ suit une loi de $\mathrm{T}^2(k, n-1)$ de Hotelling, car $\mathrm{X}_r - \mu$ suit une loi multinormale de moyenne 0 (vecteur dont les k éléments sont nuls) et de matrice de covariance Σ. De même, $\mathrm{X}^r - \mathrm{X}.$ suit une loi multinormale de moyenne 0 et de matrice de covariance $\frac{n-1}{n}\Sigma$.

Il en résulte que $\frac{n}{n-1}{}^t(\mathrm{X}_r - \mathrm{X}.)\ \mathrm{S}^{-1}(\mathrm{X}_r - \mathrm{X}.)$ suit également une loi de $\mathrm{T}^2(k, n-1)$.

Dans une représentation plane, nous n'avons que 2 dimensions. Nous en déduisons donc l'équation de la région de confiance au niveau $1 - \alpha$ de la population qui s'écrit :

$$[1] \qquad \frac{n}{n-1}[x - x.\ y - y.] \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} x - x. \\ y - y. \end{bmatrix} \leqslant t^2(\alpha, 2, n-1)$$

avec :

n : nombre d'individus de la population;
x et y : coordonnées d'un point appartenant à la région de confiance;
$x.$ et $y.$: coordonnées du barycentre de la population;
σ_x^2, σ_y^2 et σ_{xy} : variances et covariance de la population;
$t^2(\alpha, 2, n-1)$: valeur lue dans la table de la variable T^2 pour le seuil α, 2 dimensions et un nombre de degrés de liberté de $n - 1$.

Remarquons au passage que la table T^2 se déduit sans difficulté de celle du F de Fischer, car il existe entre ces deux lois la relation suivante :

$$\mathrm{T}^2(k, n-1) = \frac{k(n-1)}{n-k}\mathrm{F}(k, n-k)$$

Le développement de l'inéquation (1) conduit à une surface limitée par une ellipse d'équation paramétrique en $\Theta \in (0, 2\pi)$:

$$[2] \qquad \begin{aligned} \mathrm{X} &= x. + a\cos\gamma\cos\Theta - b\sin\gamma\sin\Theta \\ \mathrm{Y} &= y. + a\sin\gamma\cos\Theta + b\cos\gamma\sin\Theta \end{aligned}$$

où

$$\gamma = \frac{1}{2} \operatorname{arctg} \left[\frac{2\,\sigma_{xy}}{\sigma_x^2 - \sigma_y^2} \right]$$

$$a = \sqrt{\frac{n-1}{n} \frac{(\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2)\, t^2\, (\alpha,\, 2,\, n-1)}{\sigma_x^2 \sin^2 \gamma + \sigma_y^2 \cos^2 \gamma - 2\,\sigma_{xy} \sin \gamma \cos \gamma}}$$

$$b = \sqrt{\frac{n-1}{n} \frac{(\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2)\, t^2\, (\alpha,\, 2,\, n-1)}{\sigma_x^2 \cos^2 \gamma + \sigma_y^2 \sin^2 \gamma + 2\,\sigma_{xy} \sin \gamma \cos \gamma}}$$

Si $\sigma_x^2 = \sigma_y^2$, γ prend la valeur $\pi/4$ et

$$a = \sqrt{\frac{n-1}{n} (\sigma_x^2 + \sigma_{xy})\, t^2} \quad \text{et} \quad b = \sqrt{\frac{n-1}{n} (\sigma_x^2 - \sigma_{xy})\, t^2}$$

DISCUSSION

La condition essentielle pour que ce mode de représentation soit valide réside dans la binormalité des populations représentées sur les plans discriminants. Les axes discriminants sont des combinaisons linéaires des variables initiales. Ces dernières sont des moyennes de mesures morphologiques. Il apparaît donc que nous soyons dans une situation assez favorable pour admettre la normalité des variables.

Cette méthode a été développée dans le cadre de l'utilisation de moyens informatiques et prend tout son intérêt si l'on peut disposer d'une table traçante. Le jeu d'équations [2] se prête en effet sans grande difficulté à la programmation.

Un exemple de tracé témoin apparaît à la figure 1 où sont représentés à la fois les points-colonies et les ellipses de chacune des populations. Il est possible de constater la bonne coïncidence entre les nuages réels et les surfaces elliptiques ainsi qu'entre le seuil prévu (5 %) et la fraction de colonies extérieures aux ellipses (7 sur un total de 159, soit 4,4 %).

Outre l'augmentation de la lisibilité et son automatisation possible, ce type de représentation permet d'apporter une certaine objectivité dans le tracé des contours d'une population.

GRAPHIC REPRESENTATION OF MULTINORMAL POPULATIONS BY CONFIDENCE ELLIPSES

Introduction

When a new honeybee population is under a biometrical study, it is generally compared to samples of known origin. Usually (CORNUET *et al.*, 1975; GADBIN *et al.*, 1979; CORNUET *et al.*, 1982), we

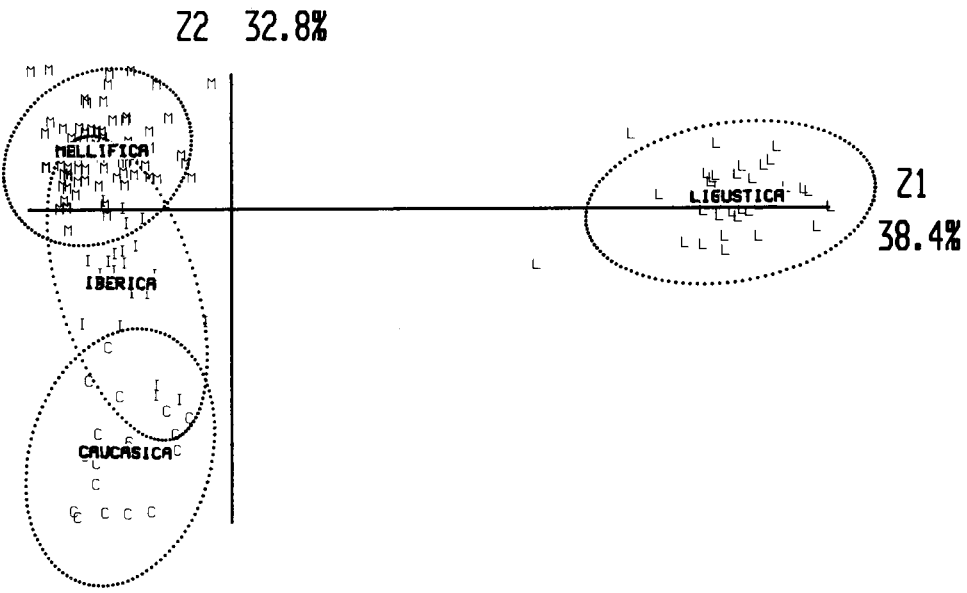

FIG. 1. — *Représentation graphique du premier plan d'une analyse factorielle discriminante réalisée sur les 4 races* mellifica, iberica, ligustica et caucasica.
Les points-colonies sont repérés par l'initiale de la race à laquelle ils appartiennent. Les ellipses de confiance sont calculées au niveau 95 %.

FIG. 1. — *Graphic representation of the first plane of a discriminant factorial analysis realized with the 4 races* mellifera, iberica, ligustica, and caucasica.
The colony-points are identified by the initial letter of their proper race. The confidence ellipses are at the 95 % level.

locate unknown colonies in the first planes of a discriminant factorial analysis carried out on the basis of reference populations. The latter form clouds of more or less distinct colony-points, which number may be quite important (generally between 100 and 200).

Consequently, the graphic representation becomes confusing and the interpretation of results, difficult. That is why, in order to simplify these representations, we thought of replacing reference colony-points by ellipses which would include statistically a given proportion of individuals, the multinormality of populations being assumed.

Rationale

Let us consider a population of n individuals characterized by k variables. With each individual r, we associate an aleatory vector X_r of dimension k. As it is classically assumed, these n aleatory vectors are supposed to be independant and to follow the same law of probability with a mean μ and a covariance matrix Σ.

Let X. be the mean vector ($X. = \frac{1}{n} \sum_r X_r$).

The aleatory matrix $S = \frac{1}{n-1} \sum_r (X_r - X.)'(X_r - X.)$ is an unbiased estimator of Σ.

If we assume that the law of probability of $X_{.r}$ is multinormal, we then know that the variable $n'(X_r - \mu)\ S^{-1}(X_r - \mu)$ follows a law of $T^2\ (k,\ n-1)$ of Hotelling, since $X_{r-}\mu$ follows a multinormal law with a mean vector O (vector which k elements are O) and a covariance matrix Σ. In the same way, $X_r - X.$ follows a multinormal law with a mean vector O and a covariance matrix $\frac{n-1}{n}\Sigma$. Hence, $\frac{n}{n-1}'(X_r - X.)\ S^{-1}\ (X_r - X.)$ follows a law of $T^2\ (k,\ n-1)$ as well.

In a plane representation, there are only two dimensions. The equation of the confidence area at the $(1-\alpha)$ level of the population is therefore as follows :

[1]
$$\frac{n}{n-1}\ [x - x.\ y - y.] \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} x - x. \\ y - y. \end{bmatrix} \leqslant t^2\ (\alpha, 2, n-1)$$

with :

n : number of individuals of the population;
$x,\ y$: coordinates of a point belonging to the confidence area;
$x.,\ y.$: coordinates of the mean of the population;
$\sigma_x^2,\ \sigma_y^2,\ \sigma_{xy}$: variances and covariance of the population;
$t^2\ (\alpha, 2, n-1)$: value read in the table of the T^2 variable for a $(1-\alpha)$ level, two dimensions and $(n-1)$ degrees of freedom.

The T^2 table is easily deduced from the F (Fischer) table, since these two laws are related as follows :

$$T^2\ (k, n-1) = \frac{k\ (n-1)}{n-k} F\ (k, n-k)$$

Resolving the inequality (1) leads to an area limitated by an ellipse which parametric equation in Θ, $(\Theta \in (0, 2\pi))$ is :

[2]
$$\begin{aligned} X &= x. + a \cos \gamma \cos \Theta - b \sin \gamma \sin \Theta \\ Y &= y. + a \sin \gamma \cos \Theta + b \cos \gamma \sin \Theta \end{aligned}$$

where :

$$\gamma = \frac{1}{2} \text{arctg}\ \frac{2\ \sigma_{xy}}{\sigma_x^2 - \sigma_y^2}$$

$$a = \sqrt{\frac{n-1}{n} \frac{(\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2)\ t^2\ (\alpha, 2, n-1)}{\sigma_x^2 \sin^2 \gamma + \sigma_y^2 \cos^2 \gamma - 2\ \sigma_{xy} \sin \gamma \cos \gamma}}$$

$$b = \sqrt{\frac{n-1}{n} \frac{(\sigma_x^2 \sigma_y^2 - \sigma_{xy}^2)\ t^2\ (\alpha, 2, n-1)}{\sigma_x^2 \cos^2 \gamma + \sigma_y^2 \sin^2 \gamma + 2\ \sigma_{xy} \sin \gamma \cos \gamma}}$$

If $\sigma_x^2 = \sigma_y^2$, we have $\gamma = \pi/4$,

$$a = \sqrt{\frac{n-1}{n} (\sigma_x^2 + \sigma_{xy})\ t^2} \qquad \text{and} \qquad b = \sqrt{\frac{b-1}{\quad} (\sigma_x^2 - \sigma_{xy})\ t^2}.$$

Discussion

This type of representation is valid only when the binormality of populations in the discriminant planes can be assumed. The discriminant axes are linear combinations of the initial variables. The latter are means of morphometric data. Therefore, we are in a fair position to admit the normality of the variables.

This method was worked out to take advantage of the use of a computerized plotter. As a matter of fact, the formulas (2) are easily programmable.

An example of a test drawing is given (Fig. 1) displaying both colonies and ellipses of each population. A good agreement is noticed between the actual clouds and the elliptical surfaces and between the given level (95 %) and the proportion of colonies within their ellipses (152/157 = 95,6 %).

In addition to the increase in legibility and the possible computerization, this type of representation gives a certain objectivity to the drawing of the outlines of populations.

Reçu pour publication en décembre 1981.
Eingegangen im Dezember 1981.

RÉFÉRENCES BIBLIOGRAPHIQUES

CORNUET J.-M., J. FRESNAYE et L. TASSENCOURT, 1975. — Discrimination et classification de populations d'abeilles à partir de caractères biométriques. *Apidologie,* **6** (2), 145-187.

CORNUET J.-M., J. ALBISETTI, N. MALLET et J. FRESNAYE, 1982. — Étude biométrique d'une population d'abeilles landaises. *Apidologie,* **13** (2), 3-13.

GADBIN C., J.-M. CORNUET et J. FRESNAYE, 1979. — Approche biométrique de la variété locale d'*Apis mellifica* L. dans le sud tchadien. *Apidologie,* **10** (2), 137-148.