

# REVUE DE STATISTIQUE APPLIQUÉE

J.-M. LEGAY

R. TOMASSONE

## **La comparaison de régressions orthogonales**

*Revue de statistique appliquée*, tome 47, n° 2 (1999), p. 81-101

[http://www.numdam.org/item?id=RSA\\_1999\\_\\_47\\_2\\_81\\_0](http://www.numdam.org/item?id=RSA_1999__47_2_81_0)

© Société française de statistique, 1999, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## LA COMPARAISON DE RÉGRESSIONS ORTHOGONALES

J.-M. Legay\*, R. Tomassone\*\*

\* Université Claude Bernard Lyon I, 69622 Villeurbanne cedex

\*\* Institut National Agronomique, 75231 Paris cedex 05

### RÉSUMÉ

La comparaison de droites de régression orthogonales est rarement présentée dans les études de statistiques appliquées. Cet article expose comment traiter la question du parallélisme d'axes principaux calculés dans des analyses en composantes principales sur plusieurs populations. Il applique les résultats à deux problèmes issus de situations biologiques particulières; en outre, il montre comment s'assurer que les premiers axes principaux sont confondus ou seulement parallèles.

*Mots-clés : analyse en composantes principales, comparaison de populations, axes parallèles, axes confondus*

### ABSTRACT

The comparison of orthogonal regression is seldom presented in applied statistics. This paper explains how we may test parallelism of principal axes computed in sample from different populations. Results are applied to two biological problems; moreover, it is shown how it is possible to test the identity of the first principal axes.

*Keywords : principal components analysis, several groups, parallel axes, confounded axes*

Dès les premiers travaux sur les corrélations et régressions (Galton, 1889), les biologistes eurent l'occasion de vouloir comparer deux ou plusieurs régressions qui représentaient, pour une même situation, des matériels différents ou des influences variées de facteurs du milieu ou bien encore des situations référées au temps ou à l'espace.

Il s'agissait d'abord de régressions où l'une des variables est privilégiée par rapport à l'autre en ce sens qu'elle constitue un repère ordonné discret (ou discrétisé) et qu'on peut l'appeler variable indépendante ou aussi variable prédictrice. Dans le cas où ces régressions sont linéaires, on connaît depuis longtemps des méthodes de comparaison (comparaison des pentes, des ordonnées à l'origine, analyse de variance, test de linéarité, etc.) (Vessereau, 1960; Zar, 1984; Tomassone *et al.* 1992).

Mais les deux variables intervenant dans une relation peuvent être des variables aléatoires classiques et donc par nature symétriques. Le besoin de comparaison est

toujours aussi important, en particulier dans beaucoup de problèmes d'allométrie, mais des méthodes complètement satisfaisantes n'étaient pas disponibles, si bien que les auteurs ont employé divers artifices, comme par exemple la discrétisation de l'une des variables, celle qu'on a quelque raison de supposer indépendante, en regroupant ou non les données en classes, ce qui permet de se ramener à l'exercice précédent.

Nous voudrions précisément, sur deux types d'exemples permettant d'aborder quelques cas possibles parmi les plus importants, décrire les procédures qui peuvent être employées pour discuter de telles relations, qu'on a appelé régressions orthogonales (le terme est de Ragnar Frisch, qui partage le premier prix Nobel d'Economie en 1969).

Le premier exemple que nous allons aborder concerne, chez le blé, la relation entre le poids de l'épi et le poids des grains de cet épi. Nous disposions d'échantillons d'un certain nombre de variétés de blé, cultivées en Tunisie et récoltées en 1947. Les échantillons de 1947 avaient donné lieu à un article (Valdeyron, Séguéla et Legay, 1949) et l'un d'entre nous avait été à l'origine de l'acquisition des premières données dans ce travail. A l'époque seul le niveau élevé des corrélations entre les deux variables avait été souligné, car il permettait de substituer au poids des grains celui de l'épi, beaucoup plus rapide à obtenir dans des perspectives de sélection agronomique (ou variétale). L'aspect régression n'avait pas été abordé, ce qui ne permettait donc pas d'interprétation dans des perspectives d'allocation de ressources.

Le deuxième exemple que nous allons analyser concerne les dimensions (longueur, largeur) d'œufs d'insectes provenant d'un travail portant sur plus de 350 espèces réparties dans 5 ordres différents. Cette fois les perspectives étaient à la fois morphogénétiques (stabilité d'une forme) et systématiques (situations d'allométrie différentes selon les niveaux taxinomiques). L'un d'entre nous avait acquis les données et commencé leur interprétation (Legay, 1977), en particulier pour tester une relation (MacMahon, 1973) avancée dans la littérature entre la longueur et la largeur (ou diamètre) d'objets biologiques, plus ou moins élastiques, comme peuvent l'être les œufs d'insectes.

Avant d'examiner ces cas concrets, nous souhaitons donner dans un premier temps les justifications formelles des procédures que nous allons employer.

## 1. Rappels théoriques : comparaison d'axes principaux

### 1.1 Analyse en composantes principales et structure de covariances

L'analyse en composantes principales (ACP) est une des méthodes de réduction de données les plus largement utilisées. Connue depuis le début du siècle, elle a été largement employée depuis qu'existent des programmes réalisant les calculs nécessaires et tout utilisateur de la statistique en dispose aisément sur son logiciel. Sous sa forme élémentaire, l'ACP consiste à transformer un ensemble de  $p$  variables aléatoires  $\mathbf{x}$ , généralement corrélées, en un autre ensemble de variables non corrélées  $\mathbf{z}$  : l'ACP est alors une méthode de *transformation de données* qui fait passer de  $\mathbf{x}$  à  $\mathbf{z}$  :

$$\mathbf{x} = [x_1, x_2, \dots, x_p]' \rightarrow \mathbf{z} = [z_1, z_2, \dots, z_p]' \quad (1)$$

Ce que l'utilisateur espère, en faisant cette transformation, est qu'il lui sera plus facile d'analyser  $\mathbf{z}$  que  $\mathbf{x}$  : parce que les nouvelles variables  $z_j (j = 1, \dots, p)$  ne sont pas corrélées et surtout parce qu'il peut n'en conserver qu'un faible nombre, les  $k (k < p)$  premières. En pratique, il dispose d'un tableau de données  $\mathbf{X}$  ( $n$  lignes,  $p$  colonnes) où la ligne  $i$  représente les valeurs de  $\mathbf{x}$  pour une *observation* (l'*unité expérimentale*) et l'ACP revient à passer de  $\mathbf{X}$  à  $\mathbf{Z}$ .  $\mathbf{X}$  est la matrice des données de base, où l'élément  $x_{ij}$  courant est la valeur de la variable  $x_j$  pour l'observation  $i$ ;  $\mathbf{Z}$  est la matrice transformée formée des éléments  $z_{ij}$ ; toutes deux sont *a priori* formées de  $n$  lignes et de  $p$  colonnes. Ce passage s'effectue en calculant :

- le vecteur moyenne des colonnes de  $\mathbf{X}$  :  $\bar{\mathbf{x}} = [\bar{x}_1 \ \bar{x}_2 \ \dots \ \bar{x}_p]'$ ,
- la matrice de variances-covariances  $\mathbf{S} = [s_{ij}]$ ; nous adopterons la notation  $s_i^2 = s_{ii}$  qui introduit le vecteur des écarts type  $\mathbf{s} = \text{diag}(\mathbf{S}) = \text{diag}(s_i)$ ;
- les valeurs propres  $l_j$  de  $\mathbf{S}$  et les vecteurs propres associés  $\mathbf{u}_j$ , sachant que toute matrice symétrique, non singulière (comme  $\mathbf{S}$ ) peut être réduite à une matrice diagonale  $\mathbf{L}$  prémultipliée et postmultipliée par une matrice orthogonale  $\mathbf{U}$  telle que :

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{L} \tag{2}$$

Les  $p$  valeurs propres  $l_j$  sont les éléments diagonaux de  $\mathbf{L}$ , et les  $p$  vecteurs  $\mathbf{u}_j$  sont les colonnes de  $\mathbf{U}$ . Nous supposons que  $l_1 > l_2 > \dots > l_p > 0$ .

L'ACP est généralement utilisée pour étudier un échantillon de taille  $n$ ; il est relativement rare qu'on se pose la question de la comparaison de  $G$  populations, chacune étant définie par un échantillon de taille  $n_g$ . Dans ce cas nous disposons de  $G$  matrices  $\mathbf{X}_g (g = 1, \dots, G)$ , et il existe toute une série de questions possibles (donc d'hypothèses statistiques) à étudier. Si nous supposons que chaque échantillon provient d'une population Normale définie par son vecteur moyenne  $\boldsymbol{\mu}_g$  et sa matrice de variances-covariances  $\boldsymbol{\Sigma}_g$  et si nous nous intéressons uniquement à la structure des relations entre les variables, il existe une hiérarchie de tests d'hypothèses :

1) Egalité des  $\boldsymbol{\Sigma}_g$  : l'hypothèse est alors

$$H_{01} : \ll \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \dots = \boldsymbol{\Sigma}_G = \boldsymbol{\Sigma} \gg$$

2) Proportionnalité des matrices  $\boldsymbol{\Sigma}_g$ .

3) Existence d'une matrice  $\boldsymbol{\Phi}$  commune de facteurs ou d'axes principaux (qui sont alors parallèles ou confondus);  $\boldsymbol{\Lambda}_g$  étant la matrice diagonale des valeurs propres du groupe  $g$  :

$$H_{03} : \ll \boldsymbol{\Phi}'\boldsymbol{\Sigma}_g\boldsymbol{\Phi} = \boldsymbol{\Lambda}_g \gg \tag{3}$$

4) Existence d'un nombre  $k (k < G)$  d'axes principaux communs.

Les ouvrages classiques de statistique multidimensionnelle traitent généralement des hypothèses 1 et 2, voir par exemple dans Kshirsagar (1972). Il existe des tests généralement introduits dans les logiciels les plus courants; souvent dans les procédures de discrimination qui imposent que la première hypothèse soit satisfaite pour pouvoir faire une analyse discriminante linéaire. Le problème qui va nous intéresser ici est celui de la troisième hypothèse : «*les axes principaux des  $G$  populations*

peuvent-ils être considérés comme identiques?». La quatrième, décrite dans Flury (1984), ne sera pas abordée.

### 1.2. Axes principaux communs (parallèles ou confondus) à plusieurs populations (Flury, 1984)

#### 1.2.1 Estimation des axes principaux communs

S'il existe une matrice  $\Phi$  commune aux  $G$  groupes, il est alors possible de calculer des matrices  $Z_g$  transformées des  $X_g$  se référant aux axes principaux communs :

$$Z_g = X_g \Phi \quad (4)$$

Notons qu'ici le point moyen du nuage de la population a pour coordonnées dans la transformation :

$$\bar{z}_g = \Phi' \bar{x}_g$$

Pour estimer  $\Phi$ , Flury (1984) calcule la *fonction de vraisemblance commune* aux  $\Sigma_g$  (où  $\Sigma_g$  est exprimée en fonction de  $\Phi$  en utilisant (3)) connaissant les  $S_g$  :

$$L(\Sigma_1, \dots, \Sigma_G) = C \prod_{g=1}^G \text{etr} \left( -\frac{n_g - 1}{2} \Sigma_g^{-1} S_g \right) |\Sigma_g|^{-(n_g - 1)/2} \quad (5)$$

où  $C$  est une constante ne dépendant pas des  $\Sigma_g$ , *etr* la fonction exponentielle de la trace et  $|\Sigma_g|$  le déterminant de  $\Sigma_g$ . En fait on se ramène à la minimisation de :

$$g(\Sigma_1, \dots, \Sigma_G) = \sum_{g=1}^G (n_g - 1) \log |\Sigma_g| + \text{tr}(\Sigma_g^{-1} S_g) \quad (6)$$

Si l'hypothèse  $H_{03}$  est vraie, on a une matrice  $\Phi$  commune, mais les valeurs propres peuvent être différentes pour chaque population  $g$ , on pose donc :

$$\Phi = [\varphi_1, \dots, \varphi_p]' \quad (7)$$

$$\Lambda_g = \text{diag}(\lambda_{g1}, \dots, \lambda_{gp}) \quad (8)$$

La minimisation conduit à résoudre les  $p(p-1)/2$  équations :

$$\varphi_h' \left( \sum_{g=1}^G (n_g - 1) \frac{\lambda_{gh} - \lambda_{gj}}{\lambda_{gh} \lambda_{gj}} S_g \right) \varphi_j = 0 \quad (h, j = 1, \dots, p, h \neq j) \quad (9)$$

avec les contraintes :

$$\Phi' \Phi = I_p \quad (10)$$

et les restrictions sur les valeurs propres :

$$\lambda_{gj} = \boldsymbol{\varphi}'_j \mathbf{S}_g \boldsymbol{\varphi}_j \quad (g = 1, G; j = 1, \dots, p) \quad (11)$$

Flury et Constantine (1985) ont publié un algorithme permettant d'effectuer ces calculs. Une fois les estimations,  $\mathbf{F}$  de  $\boldsymbol{\Phi}$  et  $\mathbf{L}_g$  des  $\boldsymbol{\Lambda}_g$ , obtenues, il est facile d'estimer  $\boldsymbol{\Sigma}_g$  par  $\mathbf{F} \mathbf{L}_g \mathbf{F}' (g = 1, \dots, G)$ . On en déduit alors la *statistique du rapport des log-vraisemblance* :

$$X^2 = \sum_{g=1}^G (n_g - 1) \log \frac{|\mathbf{F} \mathbf{L}_g \mathbf{F}'|}{|\mathbf{S}_g|} \quad (12)$$

En appliquant la théorie générale des tests de rapport de vraisemblance cette statistique, sous  $H_{03}$ , suit asymptotiquement (donc quand  $\min_{1 \leq g \leq G} (n_g) \rightarrow \infty$ ) une loi du  $\chi^2$  à  $(g - 1)p(p - 1)/2$  degrés de liberté.

### 1.2.2 Procédure simplifiée

Krzanowski (1984) a proposé une procédure de calcul simplifiée pour estimer  $\boldsymbol{\Phi}$ , elle consiste à calculer les vecteurs propres de la matrice de variances-covariances moyenne :

$$\mathbf{S} = \frac{\sum_{g=1}^G (n_g - 1) \mathbf{S}_g}{\sum_{g=1}^G (n_g - 1)} \quad (13)$$

Bien que les propriétés de cette procédure n'aient pas été complètement étudiées, elle a donné d'excellents résultats sur les exemples classiques de la littérature statistique comme, bien évidemment, les *Iris* de Fisher. Elle présente des calculs «allégés» réalisables avec les logiciels classiques.

### 1.2.3 Propriétés et nouvelle écriture de la statistique de test

On peut, en remplaçant  $\boldsymbol{\Phi}$  par  $\mathbf{F}$ , estimer les composantes principales de chacune des populations  $g$  par  $\mathbf{Z}_g = \mathbf{X}_g \mathbf{F}$ , et poser :

$$\mathbf{F}_g = \mathbf{F}' \mathbf{S}_g \mathbf{F} \quad (g = 1, \dots, G) \quad (14)$$

Puisque, d'après (11),  $\mathbf{L}_g = \text{diag}(\mathbf{F}_g)$ , il est facile d'écrire la statistique  $X^2$ , comme une fonction des seules matrices  $\mathbf{F}_g$  :

$$X^2 = \sum_{g=1}^G (n_g - 1) \log \frac{|\text{diag} \mathbf{F}_g|}{|\mathbf{F}_g|} \quad (15)$$

Le test apparaît donc comme une mesure de la capacité de *diagonaliser simultanément* les  $G$  matrices  $\mathbf{S}_g$ . La matrice  $\mathbf{F}$  définit donc une matrice de transformation qui permet d'obtenir, dans les  $G$  populations, des variables *aussi peu corrélées que possible*; la statistique  $X^2$  nous indique si cette transformation est statistiquement acceptable.

#### 1.2.4 Question complémentaire : alignement des points moyens

Une question complémentaire peut se poser si  $H_{03}$  est acceptée : «*peut-on admettre que les différents points moyens des populations sont alignés sur le premier axe principal?*». Si tel est le cas, l'utilisateur peut prendre le même axe principal pour toutes les populations. Pour répondre à cette question, il suffit de faire un test de discrimination des populations et de ne trouver qu'une seule variable discriminante : le premier axe principal. Cette discrimination se fait en utilisant les matrices de données transformées  $\mathbf{Z}_g$ .

### 1.3. Cas de deux variables

Dans le cas où  $p = 2$ , il est possible de mieux visualiser le résultat du test statistique en s'intéressant aux angles des axes principaux avec  $x_1$ , qui représente l'axe horizontal, et  $x_2$  l'axe vertical; on sait que l'ellipse de dispersion de probabilité  $100(1 - \alpha)\%$  est donnée par l'équation :

$$\frac{s_1^2 s_2^2}{s_1^2 s_2^2 - s_{12}^2} \left[ \frac{(x_1 - \bar{x}_1)^2}{s_1^2} + \frac{(x_2 - \bar{x}_2)^2}{s_2^2} - \frac{2s_{12}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{s_1^2 s_2^2} \right] = T_\alpha^2 = \frac{2(n-1)}{n-2} F_{2, n-2; \alpha} \quad (16)$$

où  $T_\alpha^2$  est la statistique de Hotelling liée au  $F_\alpha$ , la statistique de Fisher, comme il est indiqué ci-dessus. La pente du premier axe principal (la *droite de régression orthogonale*) est :

$$k_1 = \frac{s_2^2 - s_1^2 + \sqrt{(s_2^2 - s_1^2)^2 + 4s_{12}^2}}{2s_{12}} \quad (17)$$

La pente  $k_2$  du second axe principal est obtenue en remplaçant dans (17) le signe «+» devant la racine carrée par le signe «-». Mais dès lors que l'on a calculé les vecteurs propres de  $\mathbf{S}$ , on peut aussi exprimer  $k_1$  en fonction des éléments  $\nu_{ij}$  de la matrice  $\mathbf{V} = \mathbf{U}\mathbf{L}^{1/2}$ , par :

$$k_1 = \frac{\nu_{21}}{\nu_{11}} \quad (18)$$

qui ne fait intervenir que les caractéristiques du premier vecteur propre de  $S$ . Celle du second axe principal s'écrit :

$$k_2 = \frac{\nu_{22}}{\nu_{12}} = -\frac{\nu_{11}}{\nu_{21}} \quad (19)$$

La demi-longueur du premier axe de l'ellipse est  $\sqrt{l_1 T_\alpha^2}$  et celle du second  $\sqrt{l_2 T_\alpha^2}$ . L'estimation de l'angle du premier axe principal avec l'axe horizontal est  $\hat{\vartheta} = \arctan(k_1)$ . Pour obtenir un intervalle de confiance de la pente, il suffit de calculer :

$$\sin(2c) = \sqrt{\frac{l_1 l_2}{n-2}} \left[ \frac{2t_{n-2; \alpha/2}}{l_1 - l_2} \right] \quad (20)$$

où  $t_{n-2; \alpha/2}$  est un  $t$  de Student à  $n-2$  degrés de liberté au niveau  $\alpha/2$ . Les limites de confiance avec un coefficient de confiance  $100(1-\alpha)\%$  de  $k_1$  sont alors  $\tan(\hat{\vartheta} \pm c)$ . Ces résultats sont largement développés dans Jackson (1991) et, avec une formulation différente, dans Jolicoeur (1998). Ils permettent de voir que l'intervalle de confiance est d'autant plus étroit que  $l_1$  est grand par rapport à  $l_2$ , c'est-à-dire que le coefficient de corrélation entre  $x_1$  et  $x_2$  est élevé. Naturellement cette recherche n'a de sens que si l'ellipse n'est pas un cercle, c'est-à-dire si les deux valeurs propres sont différentes; ce problème est identique à celui du *test classique de sphéricité* ( $H_0 : \ll \lambda_1 = \lambda_2 \gg$ ); mais quand  $p = 2$ , il existe un test exact dont la statistique est :

$$F = \frac{(n-2)(l_1 - l_2)^2}{8l_1 l_2} \quad (21)$$

qui, sous  $H_0$ , suit une loi de Fisher  $F_{2, n-2}$ . Il faut donc, au préalable rejeter cette hypothèse pour que la suite ait un sens (James et Venables, 1980).

Toutes ces questions ont été abondamment étudiées par Jolicoeur et Mosimann (1968), Jolicoeur (1984, 1990, 1992).

## 2. Examen de deux exemples

### 2.1. Le cas des épis de blé

Les données utilisées sont celles qui correspondent à 4 variétés de blé du travail de 1947 et en outre à une variété de blé cultivée dans la région lyonnaise et récoltée en 1997 (soit 50 ans plus tard). L'ensemble des données est présenté dans l'Annexe 1.

La question posée est la comparaison de ces cinq échantillons quant à la relation poids de l'épi / poids des grains de cet épi (dans la suite relation EP/GR). Il semble justifié, après avoir présenté les statistiques élémentaires pour les cinq variétés (moyennes, variances, coefficients de corrélation) et les ellipses bidimensionnelles correspondantes, d'envisager deux étapes concernant les axes principaux :

- La première sera de tester le parallélisme des axes; si celui-ci est repoussé, la suite n'a plus de raison d'être.



• La seconde sera de savoir si les points moyens des variétés sont sur le premier axe principal ou non.

On trouvera les statistiques élémentaires dans le tableau 1, et le positionnement des ellipses bidimensionnelles dans les figures 1.1 et 1.2.

TABLEAU 1

*Matrice de variances covariances  $S_g$  des 5 populations ( $dl = n_g - 1$ ) et matrice moyenne  $S$  ( $r$  : coefficient de corrélation entre EP et GR).*

	EP	GR	$r$
$S_1 =$ ( $dl = 24$ )	0.2225000000 0.1530625000	0.1085926667	0.98470
$S_2 =$ ( $dl = 24$ )	0.0793743333 0.0541065000	0.0404543333	0.95483
$S_3 =$ ( $dl = 24$ )	1.177229333 0.868795667	0.651587667	0.99197
$S_4 =$ ( $dl = 24$ )	0.4792956667 0.3584276667	0.2697490000	0.99683
$S_5 =$ ( $dl = 29$ )	0.4158874954 0.3526833644	0.3024214816	0.99447
$S =$ ( $dl = 125$ )	0.4724985709 0.3572258685	0.2756754477	0.98979

• *Etape du parallélisme*

L'angle du premier axe principal ( $k_1$  : valeur de la tangente,  $\hat{\vartheta}$  : valeur de l'angle en degré) a été calculé dans le cas de chaque variété, et présenté avec leur intervalle de confiance.

Le test du  $\chi^2$  (15) a été mis en œuvre afin de tester le parallélisme pour les variétés 1 et 2, pour les variétés 3 et 4 et pour les variétés 4 et 5. Ces divers calculs sont donnés dans le tableau 2.

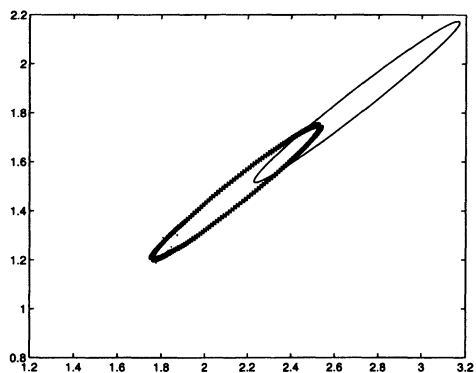


Figure 1.1. Ellipse des variétés 1 et 2, avec ellipse moyenne de ces deux variétés.  
1 : '!' ; 2 : '!' ; 1+2 : '+'

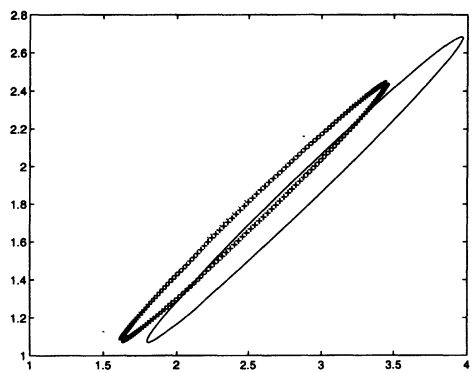


Figure 1.2. Ellipse des variétés 3 et 4, avec ellipse moyenne de ces deux variétés.  
3 : '!' ; 4 : '!' ; 3+4 : '+'

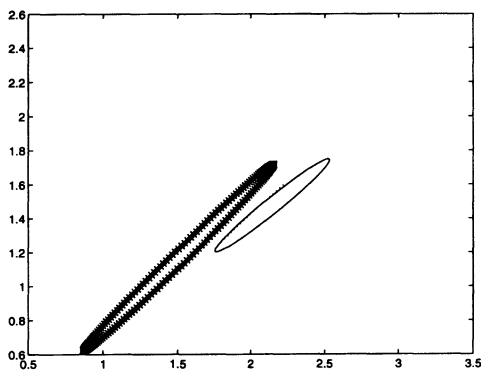


Figure 1.3. Ellipse des variétés moyenne 1+2, 3+4, et 5.  
1+2 : '!' ; 3+4 : '!' ; 5 : '\*'

FIGURE 1

Positionnement des ellipses multidimensionnelles pour les cinq variétés  
(axe horizontal : EP ; axe vertical : GR).

TABLEAU 2  
Etude du parallélisme des axes principaux

Estimation des tangentes ( $k_1$ ) et des angles ( $\hat{\vartheta}$  en degrés)  
avec leur intervalle de confiance (95%).

Pop	inf $k_1$	$k_1$	sup $k_1$	inf $\hat{\vartheta}$	$\hat{\vartheta}$	sup $\hat{\vartheta}$
1	0.6667	<b>0.6949</b>	0.7196	33.6917	<b>34.7950</b>	35.7396
2	0.6509	<b>0.7031</b>	0.7445	33.0583	<b>35.1092</b>	36.6680
1 & 2	0.6736	<b>0.6970</b>	0.7180	33.9636	<b>34.8768</b>	35.6791
3	0.7210	<b>0.7422</b>	0.7617	35.7907	<b>36.5844</b>	37.2978
4	0.7363	<b>0.7495</b>	0.7621	36.3637	<b>36.8528</b>	37.3106
3 & 4	0.7315	<b>0.7444</b>	0.7565	36.1868	<b>36.6626</b>	37.1086
5	0.8342	<b>0.8520</b>	0.8689	39.8339	<b>40.4308</b>	40.9869

Test de Flury : comparaison des cinq variétés.

<b>S</b> =	0.4725	0.3572	0.2757
<b>diag(L)</b> =	0.7446	0.0036	
<b>F</b> =	-0.7955	0.6060	-0.6060
	-0.6060	-0.7955	

<b>FL<sub>1</sub>F'</b> =	0.3282	0.0029	$r = -0.4657$
	-0.0143		
<b>FL<sub>2</sub>F'</b> =	0.1172	0.0026	$r = -0.2523$
	-0.0044		
<b>FL<sub>3</sub>F'</b> =	1.8218	0.0070	$r = -0.2003$
	-0.0226		
<b>FL<sub>4</sub>F'</b> =	0.7479	0.0011	$r = -0.1991$
	-0.0058		
<b>FL<sub>5</sub>F'</b> =	0.7142	0.0041	$r = 0.7229$
	0.0390		

$X^2 = 31.972$  à 4 dl

\*\*

TABLEAU 2 (Suite)

Test de Flury : comparaison des variétés 1 et 2

$(S_1 + S_2)/2 =$	0.1509 0.1036	0.0745
$\text{diag}(\mathbf{L}) =$	0.2231	0.0023
$\mathbf{F} =$	-0.8204 -0.5718	0.5718 -0.8204

$\mathbf{FL}_1\mathbf{F}' =$	0.3289 -0.0005	0.0022	$r = -0.0172$	$X^2 = 0.027$ à 1 dl ns
$\mathbf{FL}_2\mathbf{F}' =$	0.1174 0.0005	0.0024	$r = 0.0277$	

Test de Flury : comparaison des variétés 3 et 4

$(S_3 + S_4)/2 =$	0.8283 0.6136	0.4607
$\text{diag}(\mathbf{L}) =$	1.2850	0.0039
$\mathbf{F} =$	-0.8022 -0.59718	0.5971 -0.8022

$\mathbf{FL}_3\mathbf{F}' =$	1.8221 -0.0025	0.0067	$r = -0.0224$	$X^2 = 0.199$ à 1 dl ns
$\mathbf{FL}_4\mathbf{F}' =$	0.7479 0.0025	0.0011	$r = 0.0863$	

TABLEAU 2 (*Suite et fin*)  
Test de Flury : comparaison des variétés 4 et 5

$(24\mathbf{S}_4 + 29\mathbf{S}_5)/53 =$	0.4446 0.3553	0.2876
$\text{diag}(\mathbf{L}) =$	0.7300	0.0023
$\mathbf{F} =$	-0.7797 -0.6262	0.6262 -0.7797

$\mathbf{FL}_4\mathbf{F}' =$	0.7471 -0.0250 0.0019	$r = -0.6577$	$X^2 = 22.226$ à 1 dl ns
$\mathbf{FL}_5\mathbf{F}' =$	0.7158 0.0207 0.0025	$r = 0.4854$	

En conclusion de cette première étape, on peut dire que le test global sur l'ensemble des variétés entraîne le rejet d'axes principaux communs aux cinq variétés. Par contre on peut noter qu'il y a parallélisme entre les axes principaux des variétés 1 et 2 et entre ceux de 3 et 4. Il n'y a pas parallélisme entre les axes des variétés 4 et 5; pour ce groupe l'analyse de discrimination s'arrête là.

• *Etape de la confusion*

On va chercher à savoir maintenant si les axes sont confondus ou si, bien que parallèles, ils restent distincts.

*Cas du groupe 1 et 2*

Le test de comparaison des deux matrices n'est pas significatif puisque la valeur du  $X^2 = 0.027$  à 1 degrés de liberté (tableau 2). La structure des deux premières composantes principales est :

composante principale → Variable ↓	$z_1$	$z_2$
EP	0.8204	-0.5718
GR	0.5718	0.8204

Le test de discrimination des deux variétés n'est significatif (fortement) que pour la première composante. La seconde n'apporte aucune amélioration, car le test

de l'hypothèse « $z_2$  n'est pas significative après que  $z_1$  ait été introduite dans la discrimination», conduit à la probabilité  $P\{z_2/z_1\} = 0.3761$ , donc à l'acceptation de cette hypothèse. On peut donc admettre que le premier axe principal est *confondu pour les deux variétés*.

Les équations sont :

– pour la variété 1 :  $EP = 1.42 + 0.6949 GR$

– pour la variété 2 :  $EP = 0.81 + 0.7031 GR$

– pour l'ensemble des deux variétés :  $EP = 1.12 + 0.6970 GR$

C'est ce que symbolise le graphique des composantes principales des variétés 1 et 2 de la figure 2.

#### Cas du groupe 3 et 4

Il existe des axes principaux communs ( $X^2 = 0,199$ , tableau 2). Le test de comparaison des deux matrices n'est pas significatif. La structure des deux premières composantes principales est :

composante principale → Variable ↓	$z_1$	$z_2$
EP	0.8022	- 0.5971
GR	0.5971	0.8022

D'après le graphique des composantes principales des variétés 3 et 4 de la figure 2, la variété 3 est plus étendue que la variété 4 sans que le premier axe paraisse différent. Le test de discrimination indique que les deux variétés sont différentes, en particulier sur le facteur 2 (bien qu'elles le soient aussi sur le facteur 1). Donc les axes principaux sont parallèles, mais non confondus.

Les équations sont :

– pour la variété 3 :  $EP = 1.48 + 0.7443 GR$

– pour la variété 4 :  $EP = 0.96 + 0.7443 GR$

On ne peut pas prendre une même équation pour les deux variétés. C'est ce que traduit le graphique dans le plan des deux premières composantes des variétés 3 et 4 de la figure 2.

## 2.2 Le cas des œufs d'insectes

Cette fois nous avons choisi trois groupes, qui sont trois ordres d'insectes :

- 1) Coléoptères (parmi lesquels le hanneton, le doryphore, la coccinelle...),
- 2) Hémiptères (avec les punaises, les pucerons, les cigales...),
- 3) Hyménoptères (avec les abeilles, les fourmis, les guêpes...). L'objet étudié est l'œuf de ces insectes, caractérisé par deux variables, la longueur ( $X_1$ ) et la largeur ( $X_2$ ). En fait tous les calculs et représentations seront effectués sur les logarithmes

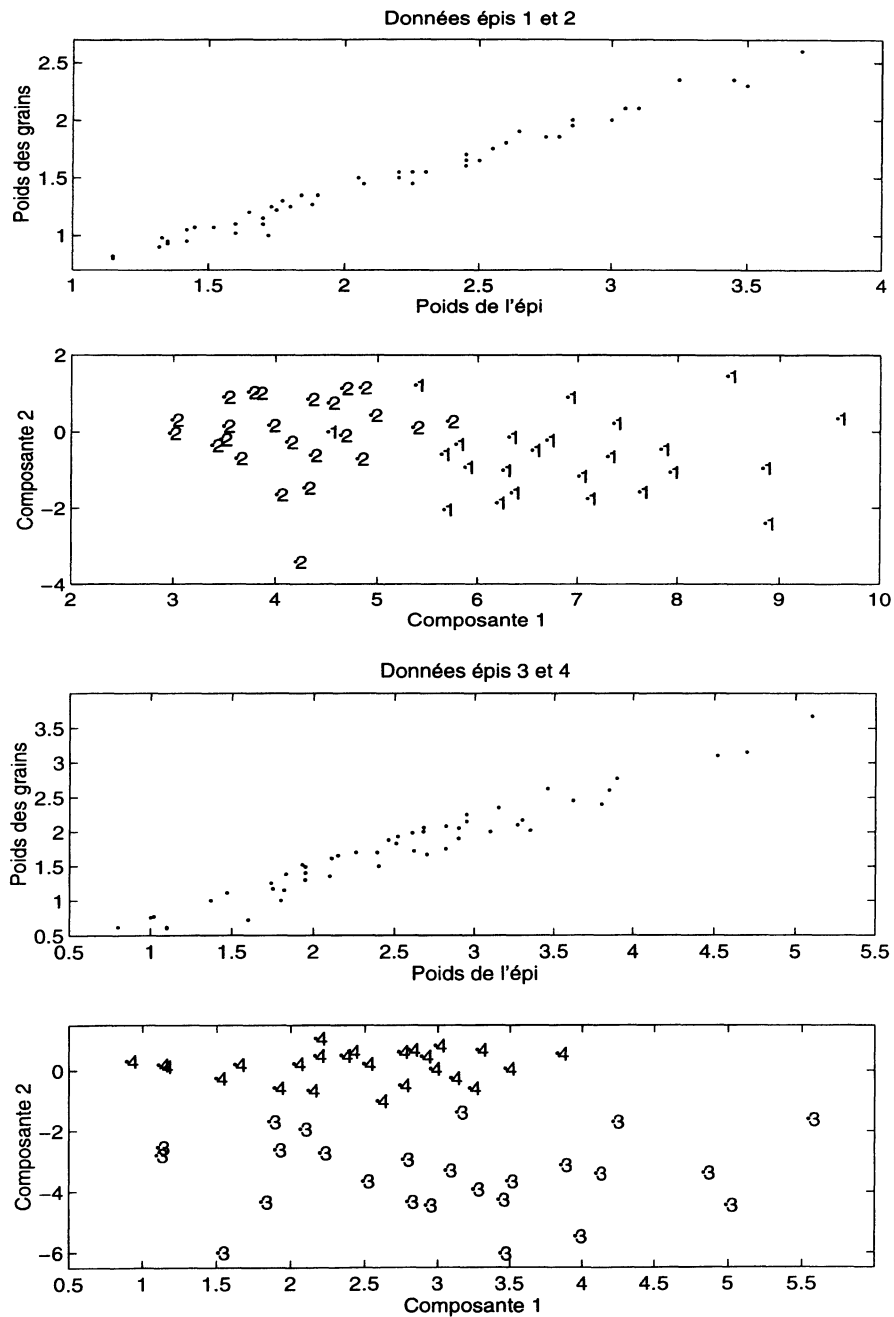


FIGURE 2  
*Comparaison des variétés 1 et 2 et des variétés 3 et 4.*

des mesures, en réalité 100 fois ces logarithmes, soit :  $L = 100 \log (X_1)$ ,  $l = 100 \log (X_2)$ . Les données sont présentées en Annexe 2. Comme précédemment nous suivrons deux étapes.

• *Etape du parallélisme*

Les trois groupes admettent-ils un même ensemble d'axes principaux? Les matrices sont données au tableau 3 et le test global présenté au tableau 4.

TABLEAU 3

Matrice de variances covariances  $S_g$  des 3 populations ( $dl = n_g - 1$ )  
et matrice moyenne  $S$  ( $r$  : coefficient de corrélation entre  $L$  et  $l$ )

	$L$	$l$	$r$
$S_1 =$ (dl = 66)	4760.944969 3570.791981	3882.371567	0.83056
$S_2 =$ (dl = 46)	2709.143441 2807.844100	4725.667122	0.78474
$S_3 =$ (dl = 52)	14418.32590 14229.65874	15526.31150	0.95105
$S =$ (dl = 164)	7247.536056 6736.435085	7810.886641	0.89533

La valeur du  $X^2$  est fortement significative, on ne peut donc pas admettre que les trois groupes aient un même ensemble de axes principaux. Le problème ultérieur n'a donc pas de sens. Par contre, on peut regarder les trois groupes deux à deux. Si on fait ces comparaisons les statistiques de test sont toujours des  $\chi^2$  à 1 dl.

Comparaison	Statistique de test	Résultat
1/2 (Coléo/Hémip)	10.9785	**
1/3 (Coléo/Hymén)	2.9736	NS
2/3 (Hémip/Hymén)	6.3681	**

• *Etude de la confusion*

La seule comparaison possible est donc la seconde : rien ne s'oppose à admettre que les groupes 1 et 3 aient des axes principaux communs (voir le troisième graphique



**TABLEAU 4**  
*Etude du parallélisme des axes principaux*  
 Test de Flury : comparaison des cinq variétés

<b>S =</b>	7247.536056	
	6736.435085	7810.886641
$10^4 * \text{diag}(\mathbf{L}) =$	1.4272	0.0787
<b>F</b>	0.6922	0.7217
	0.7217	-0.6922

$10^3 * \mathbf{FL}_1\mathbf{F}' =$	7.8710		
	0.5881	0.7723	$r = 0.2385$
$10^3 * \mathbf{FL}_2\mathbf{F}' =$	6.5649		
	-0.8901	0.8699	$r = -0.3725$
$10^4 * \mathbf{FL}_3\mathbf{F}' =$	2.9213		
	0.0041	0.0732	$r = -0.2003$

$X^2 = 10.948$  à 2dl  
 \*\*

de la figure 3). Le tableau 5 résume les résultats des calculs; il faut noter que les vecteurs colonne dans **F** sont pratiquement confondus avec les bissectrices (les deux variances sont pratiquement identiques).

En conclusion, les deux composantes principales sont nécessaires pour différencier les deux groupes : les axes principaux sont bien parallèles, mais non confondus :

– (1)  $L = -64.67 + l$

– (3)  $L = -220.18 + l$

### 3. Conclusions générales

Nous avons examiné une série de cas : celui du non parallélisme des axes, celui du parallélisme avec confusion des axes ou sans confusion.

Dans le cas des variétés de blé, on est frappé par la proximité des images de régression dans la relation EP/GR. Quelles que soient les variétés et le moment (la variété 5 s'exprime 50 ans après les autres!), la relation est remarquablement stable. Si l'on trouve des différences significatives dans certaines comparaisons soit pour la pente soit pour l'ordonnée à l'origine, c'est que l'analyse discriminante se révèle extrêmement précise, à la mesure même de la valeur très élevée des coefficients de corrélation. Sur le plan biologique on peut donc conclure que le partage des ressources

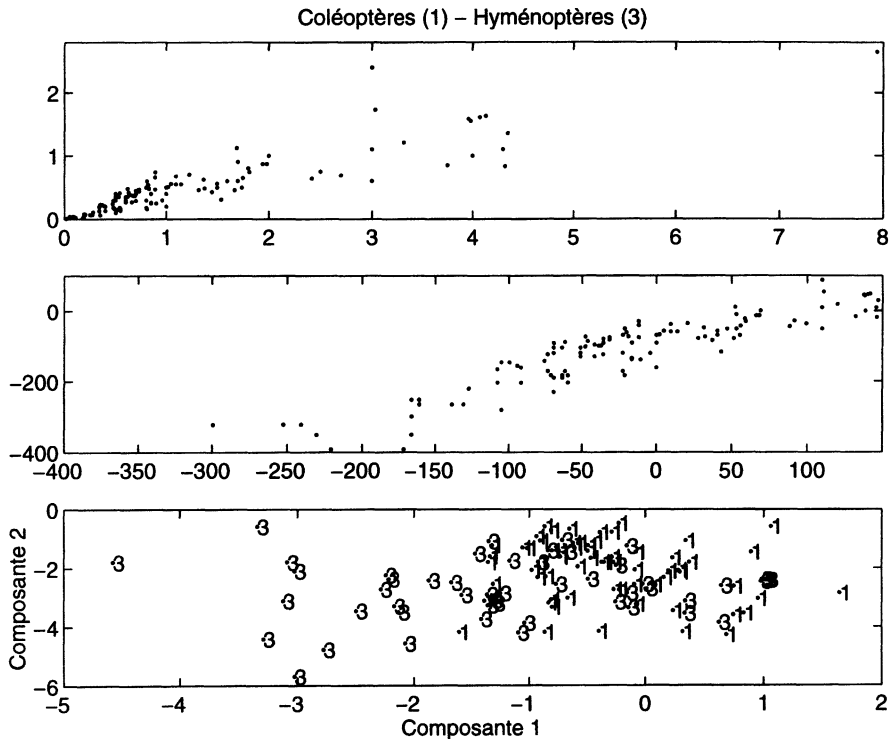


FIGURE 3

*Comparaison des Coléoptères et Hyménoptères*  
*haut : variables dans les unités d'origine*  
*milieu : variables dans les unités transformées en logarithme*  
*bas : données dans le plan des deux composantes principales.*

entre les grains et leurs enveloppes est difficilement modifiable, mais peut présenter cependant des différences identifiables.

Dans le cas des œufs d'insectes, qui ont la forme d'ellipsoïdes de révolution très allongés, on trouve aussi, dans un contexte très différent, une grande stabilité de la relation étudiée, c'est-à-dire en définitive de la forme. Mais entre *ordres* d'insectes, on constate des différences importantes soit que les pentes des axes soient différentes comme c'est le cas entre Coléoptères et Hémiptères ou entre Hémiptères et Hyménoptères, ou que ces axes soient parallèles mais nettement distincts, comme c'est le cas entre Coléoptères et Hyménoptères. Ainsi donc malgré un temps d'évolution se chiffrant par millions d'années pour les insectes, la forme de leurs œufs ne s'est pas diversifiée de façon considérable, mais des relations allométriques simples grâce à des techniques statistiques précises permettent d'identifier des situations différentes au niveau des ordres.

TABLEAU 5  
Comparaison des populations 1 et 3

## Test de Flury

$S =$	9026.3	
	8278.5	9025.1
$10^4 * \text{diag}(\mathbf{L}) =$	1.7304	0.0747
$\mathbf{F} =$	-0.7071	0.7071
	-0.7071	-0.7071

$10^3 * \mathbf{F}\mathbf{L}_1\mathbf{F}' =$	7.8925			$X^2 = 2.974$ à 1 dl ns
	-0.4390	0.7508	$r = -0.1803$	
$10^4 * \mathbf{F}\mathbf{L}_3\mathbf{F}' =$	2.9202			
	0.0555	0.0743	$r = 0.1191$	

## Discrimination sur les composantes principales communes

Palier	Composante	$F$	$(n_1, n_2)$	$P(F > F_{n_1, n_2})$	$\Lambda$ de Wilks
1	$z_1$	20.702	(1,118)	0.0001	0.85074765
2	$z_2/z_1$	11.151	(1,117)	0.0011	0.77672060
ensemble	$z_1/z_2$	16.817	(2,117)	0.0001	0.77672060

D'un point de vue méthodologique, à notre connaissance, ce type d'approche est largement méconnu, alors que les questions posées en relèvent souvent. Sans doute des améliorations sont possibles, en faisant appel en particulier à des techniques de rééchantillonnage. Comme bien souvent la stabilité des résultats, donc leur fiabilité, serait mieux étudiée; nous espérons que certains s'y attelleront.

## Références

- DAUXOIS J., ROMAIN Y., VIGUIER S. (1993), Comparison of Two Factor Subspaces. *Journal of Multivariate Analysis*, **44**, 160–178.
- FLURY B.N. (1984), Common Principal Components in  $k$  Groups. *JASA*, **79** (388), 892–898.

- FLURY B.N. (1986), Asymptotic Theory for Common Principal Component Analysis. *The Annals of Statistics*, **14** (2), 418–430.
- FLURY B.N., CONSTANTINE G. (1985), Algorithm AS211 : The F-G Diagonalization Algorithm. *Applied Statistics*, **34**, 177–183.
- GALTON F. (1889), *Natural Inheritance*. MacMillan, London.
- JACKSON J. E. (1991), *A User's guide to principal components*. Wiley, New York.
- JAMES A.T., VENABLES W. (1980), Interval estimates for bivariate principal axis, *Multivariate Analysis*, **V** (P.R. Krishnaiah, editor), North-Holland, New York, pp. 399–411.
- JOLICÉUR P., MOSIMANN J.E. (1968), Intervalles de confiance pour la pente de l'axe majeur d'une distribution normale bidimensionnelle. *Biométrie-Praximétrie*, **IX** (2), 121–140.
- JOLICÉUR P. (1984), Principal Components, Factor Analysis, and Multivariate Allometry : A Small-Sample Direction Test. *Biometrics*, **40**, 685–690.
- JOLICÉUR P. (1990), Bivariate Allometry : Interval Estimation of the Slopes of the Ordinary and Standardized Normal Major Axes and Structural Relationship. *J. Theor. Biol.*, **144**, 275–285.
- JOLICÉUR P. (1992), Bivariate allometry : Point Estimation of the Slope of the Normal Major Axis. *J. Theor. Biol.*, **154**, 35–41.
- JOLICÉUR P. (1998), *Introduction à la Biométrie*, Décarie, Montréal, 4<sup>e</sup> édition.
- KRZANOWSKI W.J. (1984), Principal component analysis in the presence of group structure. *Appl. Stat.*, **33**, 164–168.
- KSHIRSAGAR A.M. (1972), *Multivariate Analysis*. Marcel Dekker, New York.
- LEGAY J.M. (1977), Allometry and systematics Insect egg form. *J. Nat. History*, **11**, 493–499.
- MACMAHON Th. (1973), Size and Shape in Biology. *Science N.Y.*, **179**, 4079, 1201–1204.
- SCHOTT J.R. (1988), Common principal component subspaces in two groups. *Biometrika*, **75**(2), 229–236.
- SCHOTT J.R. (1991), Some tests for common principal component subspaces in several groups. *Biometrika*, **78** (4), 771–777.
- TOMASSONE R., AUDRAIN S., LESQUOY-DETURCKHEIM E., MILLIER C. (1992), *La régression : nouveaux regards sur une ancienne méthode statistique*. Masson, Paris.
- VALDEYRON G., SÉGUÉLA J., LEGAY J.M. (1949) Note sur quelques propriétés biométriques du blé et leur intérêt en matière de sélection variétale. *C.R. Acad. Agric., Fr.*, **9** (6), 1–3.
- VESSEREAU A. (1960), *Méthodes statistiques en biologie et en agronomie*. Baillière, Paris.
- ZAR J.H. (1984), *Biostatistical Analysis*. Prentice Hall, Inc., Englewood Cliffs, N.J.

**Annexe 1**

*Relation poids de l'épi (EP)/poids des grains de l'épi (GR)  
pour cinq variétés (Pop) numérotées de 1 à 5  
(de 1 à 4, Tunisie 1947, 5 Lyon 1997)*

	Pop 1 ( $n_1 = 25$ )		Pop 2 ( $n_2 = 25$ )		Pop 3 ( $n_3 = 25$ )		Pop 4 ( $n_4 = 25$ )		Pop 5 ( $n_5 = 30$ )	
	EP	GR	EP	GR	EP	GR	EP	GR	EP	GR
	3.70	2.60	1.42	0.95	3.62	2.45	2.95	2.25	0.47	0.36
	2.75	1.85	1.32	0.90	2.90	1.90	2.15	1.65	1.46	1.13
	3.25	2.35	1.80	1.25	2.10	1.35	1.95	1.40	1.39	1.06
	2.65	1.90	2.20	1.55	2.62	1.72	2.52	1.93	1.04	0.80
	3.00	2.00	1.72	1.00	3.10	2.00	2.95	2.15	1.82	1.46
	3.05	2.10	1.15	0.82	2.70	1.67	3.15	2.35	2.05	1.65
	2.30	1.55	1.35	0.93	3.85	2.60	2.68	2.00	1.19	0.78
	3.45	2.35	1.60	1.02	2.40	1.50	2.68	2.06	1.13	0.82
	2.85	2.00	1.60	1.10	4.52	3.10	2.61	1.98	1.68	1.24
	2.25	1.55	1.88	1.27	5.10	3.67	1.95	1.49	1.32	1.08
	2.25	1.45	1.45	1.07	4.70	3.15	3.46	2.62	1.61	1.32
	2.50	1.65	1.70	1.10	3.90	2.77	2.82	2.08	2.61	2.11
	3.50	2.30	1.65	1.20	1.82	1.15	2.11	1.61	0.62	0.44
	2.05	1.50	1.77	1.30	3.80	2.40	2.51	1.83	1.47	1.18
	2.45	1.65	1.70	1.15	3.27	2.10	1.00	0.76	2.23	1.69
	2.60	1.80	2.07	1.45	1.60	0.72	1.83	1.38	1.10	0.83
	2.20	1.50	1.33	0.98	1.95	1.30	2.39	1.70	1.09	0.87
	2.55	1.75	1.73	1.25	3.35	2.02	2.26	1.70	2.30	1.89
	1.75	1.22	1.84	1.35	2.82	1.75	2.46	1.88	1.41	1.11
	2.80	1.85	1.52	1.07	1.80	1.00	1.37	1.00	0.92	0.69
	2.85	1.95	1.35	0.95	2.90	2.05	1.93	1.52	0.89	0.63
	2.75	1.85	1.15	0.80	3.30	2.17	1.74	1.25	1.36	0.92
	2.45	1.60	1.42	1.05	1.10	0.62	1.47	1.11	2.17	1.64
	2.45	1.70	1.15	0.82	1.10	0.60	1.02	0.77	2.78	2.23
	3.10	2.10	1.90	1.35	1.75	1.17	0.80	0.62	1.74	1.33
									0.85	0.64
									1.71	1.39
									1.40	1.11
									3.02	2.49
									0.55	0.23
moyenne	2.70	1.84	1.59	1.11	2.88	1.88	2.19	1.64	1.51	1.17
écart type	0.47	0.33	0.28	0.20	1.09	0.81	0.69	0.52	0.64	0.55

*Echantillons réunis ( $n = 130$ )*

	EP	GR
moyenne	2.15	1.51
écart type	0.88	0.61

**Annexe 2**  
*Données insectes*

Pop : 1 : Coléoptères; 2 : Hémiptères; 3 : Hyménoptères

	Pop 1				Pop 2		Pop 3		
	X <sub>1</sub>	X <sub>2</sub>	X <sub>1</sub>	X <sub>2</sub>	X <sub>1</sub>	X <sub>2</sub>	X <sub>1</sub>	X <sub>2</sub>	
	0.70	0.29	0.81	0.60	0.84	0.60	0.11	0.02	
	2.42	0.64	0.53	0.35	1.30	0.40	0.08	0.04	
	1.69	1.12	3.00	2.40	1.67	1.01	1.70	0.60	
	0.50	0.30	1.54	0.31	0.55	0.24	0.35	0.06	
	0.48	0.29	1.50	0.56	0.57	0.33	0.09	0.04	
	1.94	0.87	0.60	0.30	2.02	0.73	1.50	0.50	
	0.62	0.48	1.80	0.80	1.08	0.22	2.50	0.75	
	0.80	0.50	0.39	0.21	0.60	0.20	3.98	1.55	
	0.82	0.54	0.37	0.23	0.57	0.27	4.13	1.63	
	0.50	0.35	1.37	0.62	1.05	0.45	3.96	1.58	
	1.02	0.51	0.68	0.37	0.84	0.51	3.00	1.10	
	0.50	0.10	0.73	0.44	0.66	0.19	4.07	1.61	
	0.48	0.18	0.85	0.40	1.34	0.93	0.80	0.18	
	0.50	0.40	0.89	0.66	0.82	0.25	0.81	0.16	
	0.60	0.27			1.05	0.22	0.53	0.16	
	1.05	0.56			0.34	0.17	0.20	0.08	
	4.35	1.35			0.86	0.20	1.10	0.68	
	0.96	0.30			0.89	0.68	0.40	0.13	
	0.85	0.26			1.14	0.92	3.75	0.85	
	1.81	0.74			0.92	0.76	0.05	0.04	
	2.00	1.00			0.80	0.30	0.18	0.02	
	0.69	0.37			0.70	0.25	0.34	0.13	
	1.74	0.50			0.49	0.23	0.55	0.16	
	1.00	0.20			0.61	0.28	0.54	0.18	
	0.62	0.36			0.86	0.65	0.40	0.13	
	0.66	0.38			1.41	0.63	0.19	0.05	
	1.75	0.65			0.41	0.13	0.80	0.30	
	4.32	0.83			0.91	0.16	0.60	0.30	
	1.70	0.90			1.05	0.89	0.60	0.35	
	0.54	0.41			1.04	0.86	0.19	0.03	
	0.90	0.25			0.55	0.25	1.00	0.40	
	4.00	1.00			1.09	0.27	1.67	0.46	
	0.83	0.48			0.30	0.17	0.63	0.42	
	0.70	0.45			0.40	0.16	1.60	0.60	
	7.94	2.65			0.31	0.12	1.70	0.60	
	1.23	0.70			0.43	0.17	0.34	0.19	
	0.66	0.27			0.44	0.12	0.47	0.24	
	0.89	0.47			0.36	0.13	0.02	0.01	
	0.89	0.74			1.50	0.80	0.49	0.16	
	1.98	0.87			1.30	0.30	1.00	0.40	
	0.40	0.20			1.38	1.22	0.70	0.40	
	0.50	0.15			1.30	1.03	1.15	0.55	
	1.38	0.48			0.50	0.33	0.10	0.03	
	1.32	0.46			0.27	0.16	0.28	0.11	
	3.00	0.60			0.67	0.26	0.53	0.15	
	4.30	1.10			2.00	0.75	0.55	0.13	
	0.85	0.25			1.39	0.53	2.70	0.69	
	1.00	0.50					0.20	0.07	
	3.03	1.73					0.27	0.07	
	1.10	0.55					0.35	0.23	
	3.32	1.21					0.25	0.07	
	1.00	0.50					1.45	0.43	
	0.73	0.47					0.19	0.08	
moyenne	1.412	0.597			0.885	0.435	1.040	0.375	unité :
écart type	1.295	0.459			0.437	0.303	1.170	0.429	origine
moyenne	7.47	- 72.14			- 24.81	- 106.51	- 57.37	- 162.81	unité :
écart type	69.00	62.31			52.05	68.74	120.08	124.60	100*log