REVUE DE STATISTIQUE APPLIQUÉE

E. MATHIEU
M. AUTEM
M. ROUX
F. BONHOMME

Épreuves de validation dans l'analyse de structures génétiques multivariées : comment tester l'équilibre panmictique ?

Revue de statistique appliquée, tome 38, n° 1 (1990), p. 47-66
<http://www.numdam.org/item?id=RSA_1990__38_1_47_0>

Rev. Statistique Appliquée, 1990, **XXXVIII** (1), 47-66

ÉPREUVES DE VALIDATION DANS L'ANALYSE DE STRUCTURES GÉNÉTIQUES MULTIVARIÉES : COMMENT TESTER L'ÉQUILIBRE PANMICTIQUE ?

E. MATHIEU*, M. AUTEM*, M. ROUX** et F. BONHOMME*

** : Institut des Sciences de l'Evolution, C.N.R.S. U.A. 327, Université de Montpellier, Place E. Bataillon - F. 34060*
*** : C.E.F.E., C.N.R.S. Route de Mende, B.P. 5051 - F. MONTPELLIER*

RÉSUMÉ

Cet article a trait aux méthodes employées en génétique des populations pour analyser les données du polymorphisme interindividuel multilocus. Après une description de la structure des données et une présentation des problèmes de significativité posés par l'interprétation des sorties des analyses factorielles les plus couramment employées, nous explorons plusieurs techniques de validation statistique.

Grâce à une hypothèse nulle réaliste (l'équilibre panmictique), nous proposons une procédure de rééchantillonnage assortie de tests de validation obtenus par l'étude des distributions de plusieurs indices de comparaison inter-tableaux. Après un test comparatif sur cas réels des diverses approches proposées, nous retiendrons celles qui nous semblent donner les meilleurs résultats et concluons sur les limitations de ces démarches.

Mots clefs : *Génétique, Analyse Multivariée, Validation, Panmixie.*

ABSTRACT

This article deals with the methods used in population genetics forthe study of multilocus interindividual polymorphism. After a short description of the data structure and of the problems of significance linked with the interpretation of the outputs yielded by the most commonly used multivariate techniques, we explore several validation procedures.

Affording a realistic nul hypothesis (the panmictic equilibrium), we proprose a sample reuse-method coupled with validation tests based on the study of statistical distribution of various inter-table indexes. Following a comparaison of the various approaches on a real cases, we evaluate their merits and conclude on their limits.

1. Introduction

Dans de nombreuses disciplines, les données disponibles pour l'analyse se présentent sous la forme de tableaux bruts (matrices rectangulaires) Objets X Descripteurs. La génétique des populations ne fait pas exception à cette règle puisque son objet est l'analyse de la diversité existant entre individus décrits par leurs gènes. Aussi a-t-on largement recours à des techniques multivariées pour extraire l'information contenue dans les données du polymorphisme inter-individuel ou inter-populationnel.

L'objet de ce travail est d'établir de manière empirique des procédures de validation statistique proposées comme aides à l'interprétation objectives des résultats d'analyses factorielles.

Avant de préciser les différentes approches abordées, il est tout d'abord nécessaire de faire quelques rappels sur la nature et la composition de l'information génétique.

1.1. L'information génétique

La continuité à travers les générations des caractères spécifiquesd'une espèce est régie par la transmission d'une information, dont Mendel a démontré la nature quantique.

Chaque unité informationnelle ou **gène** est impliquée dans l'expression d'un caractère par l'intermédiaire d'une fonction cellulaire. Ces gènes disposés le long des chromosomes, occupent chacun une position déterminée, ou **locus** du gène. Du fait de l'action des mutations, ces gènes peuvent être modifiés ce qui entraîne, dans une certaine proportion des cas, un changement du caractère qu'ils contrôlent. Ces nouveaux états possibles d'un même gène sont appelés **allèles**. S'il existe plusieurs allèles à un locus donné, il y a alors **polymorphisme** et ce polymorphisme constitue la base des observations en génétique des populations. Il y a en général plusieurs chromosomes dans une espèce, et plusieurs dizaines de milliers de gènes répartis sur ces chromosomes. Si chaque gène peut exister sous plusieurs formes allèliques, cela représente donc un nombre très grand de combinaisons possibles d'allèles. On appelle **génotype** d'un individu la combinaison d'allèles portée par celui-ci sur l'ensemble des locus analysés. Chez la plupart des espèces évoluées, les individus disposent d'un double jeu de chromosomes, l'un d'origine paternelle, l'autre d'origine maternelle. Ils sont dits **diploïdes**, c'est-à-dire que chaque gène est présent en double copie. Dans un individu, si ces allèles paternels et maternels sont identiques pour un locus donné, l'organisme est dit **homozygote** à ce locus. Dans le cas contraire, l'organisme est dit **hétérozygote** pour le locus concerné et chaque copie peut être légèrement différente de l'autre.

La génétique sait maintenant caractériser toujours plus finement le contenu informationnel des gènes, soit indirectement par analyse du produit de leur activité (les **protéines**), soit directement par analyse du support lui-même, l'**ADN**, au niveau de l'enchaînement de nucléotides qui le composent. Les analyses populationnelles

de routine permettent de caractériser l'état allèlique de quelques dizaines de locus chez plusieurs centaines, voire milliers d'individus.

1.2. Les données

1.2.1. Leur structure

Dans la nature, une espèce biologique constitue rarement un continuum homogène. Si on étudie la répartition des gènes sur l'ensemble de l'aire géographique de l'espèce, des variations locales apparaissent. Ces variations constituent le plus souvent des unités géographiques naturelles au sein desquelles les individus ont tendance à se croiser de façon préférentielle. Par hypothèse, les communautés reproductrices d'individus partageant le même "**pool**" génique ont reçu le nom de **population** (DOBZHANSKY 1955). L'existence de populations conspécifiques implique un certain degré d'isolement reproducteur mais aussi des échanges avec les populations voisines empêchant ainsi le morcellement de l'espèce en plusieurs espèces nouvelles. Dans le cadre que se fixe la génétique des populations, l'analyse du polymorphisme est un bon moyen pour aborder les structures d'échanges et de reproduction à l'intérieur des populations. Le polymorphisme génétique, notamment celui détectable par électrophorèse des protéines, recèle des informations pertinentes quant à la structure génétique de l'espèce ou plus précisément de ses échanges interpopulationnels. Une population est dite à l'équilibre **panmictique** si, lors de la reproduction sexuée, l'union des gamètes formant les oeufs (ou zygotes) est complètement aléatoire, chaque gamète emportant un allèle donné en fonction de sa seule fréquence dans la population, et aucun allèle n'étant particulièrement associé à aucun autre. Ceci se traduit non seulement par l'indépendance des allèles d'un même gène (loi de Hardy-Weinberg) mais aussi par l'absence d'associations entre gènes situés à différents locus du chromosome. Ces associations constitueraient ce qu'il est convenu d'appeler des déséquilibres gamétiques. La notion d'équilibre panmictique constitue l'hypothèse nulle des généticiens. Elle se traduit par une matrice des données brutes où les allèles sont distribués de manière aléatoire dans les individus. Mais est-on capable de mesurer des écarts à cette situation (structuration génétique).

Notons ici qu'il existe de nombreux travaux ayant pour objet de tester l'équilibre panmictique pour un seul locus à la fois (écart à la loi de Hardy-Weinberg). On consultera par exemple le travail récent de HERNANDEZ et WEIR (1989). Cependant notre propos est de prendre en compte simultanément l'information contenue dans l'ensemble des locus polymorphes. L'approche non paramétrique par rééchantillonnages successifs que nous développons nous a semblé bien adaptée.

Que se passe-t-il dans les cas concrets? On ne connaît en général de la population qu'un échantillon et le problème posé est de détecter des hétérogénéités à l'intérieur de cet échantillon. Or, si une collection d'individus est "homogène" (tirée d'une unité à l'équilibre panmictique par exemple), on s'attend à obtenir un ensemble de données brutes (matrice Individus X Gènes) "aléatoire". Si au contraire elle recèle des immigrants de groupes génétiquement différenciés (en

termes de fréquences alléliques), on s'attend alors à des distorsions entre la distribution observée des données brutes et l'attendu stochastique, et finalement à des correspondances partielles entre les allèles des gènes impliqués dans les différenciations des groupes concernés ("déficits d'hétérozygotes" et "déséquilibres gamétiques"). D'autres causes ayant également pour effet de structurer la masse des données brutes peuvent intervenir au niveau des descripteurs (gènes) dans la mesure où des associations entre ces marqueurs peuvent être le reflet de l'histoire naturelle d'une population modelée au fil des générations par tous les paramètres de son évolution (régime de reproduction, sélection différentielle des individus due aux conditions du milieu, dérive aléatoire des fréquences génotypiques due à des effectifs reproducteurs faibles ...). Ces structurations sont-elles détectables et identifiables sur une base statistique sûre?. Notre objectif étant précisément de décrire l'hétérogénéité éventuelle de la collection d'individus extraite d'une population, il n'est pas souhaitable de décrire l'organisation populationnelle à partir des seules fréquences alléliques au risque de perdre toute l'information que recèle la combinatoire des gènes.

1.2.2. Leur analyse

Le problème que nous venons d'esquisser est d'abord du ressort de l'analyse exploratoire, c'est-à-dire de la mise en évidence de structures dans la masse de données brutes. Dans notre cas, nous avons utilisé essentiellement l'Analyse Factorielle des Correspondances Multiples (AFCM). Cette technique multivariée est particulièrement bien adaptée aux données du polymorphisme génétique (AUTEM et Coll. 1987, SHE et Coll. 1987) car elle appréhende directement les corrélations entre descripteurs, c'est-à-dire les déséquilibres gamétiques, et qu'elle permet "ipso facto" de s'affranchir des épineux problèmes de pondération (BENZECRI et Coll. 1973). L'approche factorielle, qui vise à dégager sous forme "d'associations" partielles les principales structures des matrices de données, permet l'exploitation simultanée de l'information contenue dans les correspondances allèliques intralocus (écarts à l'équilibre de Hardy Weinberg) et celle recélée dans les correspondances allèliques interlocus (déséquilibres gamétiques). Les données brutes d'une collection d'individus constituent un tableau où chaque individu (en ligne) est décrit par les génotypes (en colonnes) des N locus polymorphes retenus. Il est à noter que le passage du phénotype au génotype est facilité dans notre approche grâce à la technique de caractérisation employée (électrophorèse enzymatique) qui révèle des allèles codominants. Un exemple de tableau réel est donné dans le Tableau Ia. Ce corpus de données brutes est ensuite recodé, pour les besoins de l'AFC, sous la forme d'un tableau comportant autant de colonnes que d'allèles. Chaque individu est représenté par une suite de 0, 1 ou 2. La valeur 2 indique si l'individu est homozygote pour l'allèle concerné, la valeur 1 s'il le possède en condition hétérozygote et la valeur 0 s'il ne le possède pas. (Tableau Ib).

Nos premiers résultats (c.f. exemples détaillés en § 2.4.) permettent d'entrevoir pour certains échantillons une structuration de leur variabilité (Exemples des figures 1a et 1b). Nous pensons que la mise en évidence, par analyse multivariée, d'associations entre descripteurs (allèles) ou entre objets (individus) permettra d'attribuer ces dernières à des phénomènes précis de ces organismes (migrations, sélection, dérive...) que nous avons déjà évoqués.

1.3. Le problème posé : le rétrodiagnostic en analyse multivariée

Les questions mentionnées plus haut débouchent naturellement, pour le généticien sur des problèmes de **validation statistique**, ce qui constitue l'objet de notre article :

- Les structures observées sont-elles significatives et stables?
- Peut-on s'affranchir des fluctuations d'échantillonnage?
- Comment déterminer la validité d'un résultat, et améliorer les estimations de fiabilité de nos conclusions?

L'analyse **exploratoire**, purement descriptive, qu'est l'AFC doit s'accompagner d'analyses **confirmatoires** de stabilité, c'est-à-dire d'un "**rétrodiagnostic**" comme l'appellent TUCKEY et MALLOWS (1982).

L'application du "rétrodiagnostic" permet d'employer, selon le but poursuivi, un ou plusieurs critères de qualité pour décider qu'une structure est satisfaisante. GREENACRE (1983) distingue deux types de critères de qualité concernant la stabilité d'une analyse multivariée.

La stabilité interne : Le nuage de points projeté sur le plan ne doit pas être déterminé par des aspects isolés de la collection des données (tel un point aberrant isolé qui tirerait de façon excessive le plan principal dans sa direction).

La stabilité externe : L'orientation du plan doit être peu affectée par la considération de nouveaux échantillons de la même population; un échantillon insuffisamment étendu pour pouvoir caractériser avec peu de variabilité les structures de la population est un exemple d'instabilité externe, d'autres échantillons de même taille menant à des plans principaux différents. La notion de stabilité externe couvre donc celle de la répétitivité des structures décrites.

Plusieurs outils ont été proposés à des fins d'analyses confirmatoires (HOLMES JUNCA 1985) :

– La "**validation croisée**" (STONE 1974) qui consiste à analyser la stabilité des résultats obtenus sur des partitions d'un très gros fichier de données.

– Le "**bootstrap**" (EFRON 1979, 1982) ou méthode dite "à la Cyrano" qui opère sur des rééchantillonnages avec remise à partir des données réelles.

– Le "**jackknife**" (MILLER 1974) ou méthode dite "passe-partout" où l'on décide d'omettre une variable ou un individu de la matrice de données originelles; on fabrique ainsi autant de sous échantillons qu'il y a d'individus, ou de variables.

Cependant, nous n'utiliserons aucun de ces outils car nous avons la chance de disposer d'une hypothèse nulle réaliste qui peut être approchée de manière analytique, ce qui est suffisamment rare pour mériter d'être souligné. Plaçons-nous dans notre situation hypothétique d'une population à l'**équilibre panmictique** (c.f.§ 1.2.1.). L'outil rétrodiagnostique que nous nous proposons d'utiliser repose sur cette **hypothèse nulle d'homogénéité génétique** de la population dont sont issus nos échantillons. Nous engendrons, sous cette hypothèse, des échantillons simulés qui serviront au "rétrodiagnostic". Plusieurs niveaux de discussion sont à prendre

en considération : Le premier se rapporte à la façon d'engendrer nos échantillons simulés. Le second concerne les techniques de comparaison des résultats obtenus à partir des données simulées aux résultats fournis par les données réelles. Enfin nous discuterons de l'évaluation des risques d'erreurs liés aux tests ainsi créés.

2. Matériel et méthodes

2.1. Construction d'échantillons simulés

La première solution envisagée a consisté, à partir des données réelles, à estimer les fréquences alléliques en chacun des locus étudiés. Nous disposons ainsi d'une *urne gamétique* à partir de laquelle, et selon un modèle d'union aléatoire des gamètes, on tire aléatoirement autant d'échantillons qu'on le désire; les effectifs sont égaux à celui de l'échantillon réel. Les conditions du tirage aléatoire sont les suivantes :

- Chaque allèle est tiré avec une probabilité égale à sa fréquence dans l'échantillon réel.
- Les tirages en deux locus différents sont indépendants.
- Les tirages de deux individus différents sont indépendants.

Les résultats de ces tirages, ou *simulons*, présentent deux facteurs de variabilité :

- Le premier est constitué par la variance sur les fréquences des modalités et est ce qu'on pourrait appeler la variance d'échantillonnage sur les fréquences de l'urne; puisqu'une "urne réduite" s'écarte plus ou moins de "l'urne vraie" de la population (variance d'échantillonnage *sensu stricto*).
- Le second constitue la variance sur la combinatoire multilocus engendrée sous l'hypothèse nulle d'homogénéité génétique de nos échantillons.

Ce que nous pouvons reprocher à un simulon, c'est sa faible aptitude à pouvoir représenter convenablement l'homologue stochastique de l'échantillon réel. Cela est du à la petitesse des effectifs des échantillons (de l'ordre de la centaine) relativement aux fréquences de certains allèles rares (quelques %) : Si un allèle est présent à faible fréquence dans l'échantillon réel, il peut fort bien ne pas être représenté du tout dans le tirage gamétique aléatoire. Les fréquences rares étant favorisées dans des analyses de type AFC (GAUCH 1982), les simulons apparaissent en général aussi différents entre eux que de l'échantillon réel. C'est ce que nous avons constaté dans un premier temps en comparant visuellement (décompte de points hors enveloppe) des superpositions d'images factorielles (résultats non présentés). Ces techniques n'ayant pas donné de résultats informatifs, nous avons dans la pratique préféré engendrer des tableaux de données simulées en effectuant, pour chaque locus, une permutation aléatoire des allèles présents dans l'échantillon réel. En outre les permutations relatives à deux locus

différents sont distinctes et indépendantes. En d'autres termes, on pratique un tirage gamètique sans remise, dans l'échantillon réel, jusqu'à épuisement de l'urne. A ces échantillons permutés nous avons donné le nom de "**permutons**" avec lesquels nous avons réalisé la plupart des tests développés ci-dessous. Les associations résiduelles visibles dans les permutons, ne résultent que de combinaisons aléatoires de variables puisque la variabilité liée aux fréquences est identique à celle de l'échantillon réel.

2.2. Validation statistique

Pour ce faire, nous avons essayé plusieurs approches "rétrodiagnostiques" empiriques qui nous ont fourni des résultats plus ou moins significatifs selon les métriques employées et le choix des échantillons réels de référence.

Ayant décidé d'abandonner les procédures de validation statistique basées sur l'interprétation d'images multivariées obtenues à partir d'échantillons simulés pour les raisons données plus haut, nous nous sommes orientés vers d'autres techniques non factorielles, basées sur la comparaison d'indices globaux obtenus à partir des permutons. Ces permutons sont considérés comme des réalisations particulières de populations à l'équilibre panmictique auxquelles est comparée la collection réelle d'individus. Cette comparaison s'effectue par l'étude de lois de statistiques d'homogénéité sur ces tableaux perturbés.

Parmi ces approches, les méthodes STATIS (LAVIT 1988) permet une analyse simultanée de plusieurs tableaux de données quantitatives Individus X Variables à partir des coefficients "RV" (ESCOUFIER 1973). Les tableaux sur lesquels nous avons travaillé, issus des permutons et de l'échantillon réel, sont ceux des coordonnées de tous les allèles fournies par analyses factorielles des correspondances. Cette technique qui essaie de comparer, à travers les coordonnées factorielles, les informations multidimensionnelles des nuages n'a pas non plus donné les résultats escomptés (les échantillons réels sont toujours dans la fourchette des échantillons permutés - résultats non présentés).

L'autre approche, celle que nous développerons dans le cadre de cetarticle, consiste à établir la loi de distribution des moyennes, et des variances d'indices particuliers, calculées à partir de tous les individus sur chaque permuton et pour l'échantillon réel. Les valeurs de la moyenne et de la variance du cas réel seront placées sur les lois respectives des permutons et permettront de situer cet échantillon parmi ses homologues stochastiques.

Trois indices ont été employés, les deux premiers (Degré d'homologie et identité génétique) ayant les propriétés d'une distance, il nous a été difficile de comparer leurs moyennes et variances obtenues sur permutons à celles du cas réel à l'aide des tests paramétriques usuels; les lois de distribution des moyennes et variances de ces indices sont inconnues et les différentes distances entre individus dans un tableau n'étant pas indépendantes les unes des autres, la connaissance du degré de liberté du système est loin d'être évidente. Quant au troisième indice, nous verrons plus loin qu'il est de nature probabiliste. Nous avons choisi d'engendrer pour chaque test 100 permutons servant à la construction des différentes lois, ce chiffre semble nécessaire et suffisant pour ce genre de travaux. Pour exemple, la loi

de distribution des moyennes du troisième indice est donnée figure 3. La difficulté principale réside dans l'interprétation de la position du point réel sur la loi, notion appréciative relevant du risque de deuxième espèce. Afin dévaluer et d'appréhender l'efficacité du test, nous avons réalisé plusieurs expériences sur des échantillons réels présentant *a priori* des degrés divers de structuration génétique.

2.3. Algorithme et programmes de simulation

Tous les programmes utilisés ont été écrits en FORTRAN 77 sur IBM PC/AT2 et compilés à l'aide du compilateur MICROSOFT Version 3.31 ou du compilateur FORTRAN IBM Version 1.0 (Mac FARLAND).

– *Programme générateur d'échantillons permutés*

Ce programme fonctionne de la manière suivante : En entrée, on dispose d'un tableau de données réelles Individus X Génotypes. Afin d'engendrer des échantillons simulés par permutations alléliques, il nous faut un générateur de nombres "pseudo-aléatoires" uniformément distribués sur l'intervalle [0,1]. Le processus choisi comme générateur de nombres aléatoires (distribués selon une loi continue uniforme sur [0,1]) repose sur la méthode dite congruentielle multiplicative (LEHMER 1951). Notre souci étant d'obtenir une bonne qualité statistique de notre suite "pseudo-aléatoire" nous avons utilisé la méthode du brassage de deux générateurs. L'intérêt de cette technique tient à la qualité du générateur résultant, qui est toujours bonne, même si les deux générateurs de base sont relativement médiocres, et à la période obtenue qui est le produit des deux périodes (Dans notre cas, pour une machine à mots de 16 bits la période résultante est : $T = T1.T2 = 2^{16}.2^{16} = 2^{32}$).

– *Programmes de calcul d'indices*

Ces programmes utilisent en entrée les différents tableaux permutés engendrés précédemment et le tableau réel, soit dans leur forme brute, soit dans leur forme codée disjonctive. Ces programmes travaillent en double précision et sont conçus pour traiter plus de 100 tableaux d'affilée.

– *Programmes d'analyses factorielles des correspondances*

Les tableaux obtenus et préalablement codés peuvent être soumis sur micro-ordinateur au programme d'Analyse Factorielle des Correspondances élaboré par Maurice ROUX (1988) et inclus dans le progiciel BIOMECO (C.E.P.E., Montpellier).

2.4. Echantillons tests

Ce sont ceux présentés en figure 1a, 1b et en figure 2.

La figure 1 représente les projections des points-lignes (individus) et des points colonnes (allèles) de deux espèces voisines de rongeurs, décrites par les mêmes locus (gènes) et analysées par A.F.C.. Nous avons superposé à ces projections les enveloppes convexes représentant les provenances géographiques des individus (d'après J.M. DUPLANTIER, Thèse 1988). De manière très claire, l'échantillon "B" (deuxième espèce) présente un allongement significatif sur l'axe I et un début de différenciation géographique alors que ce phénomène n'est pas du tout apparent dans l'échantillon "A" (première espèce). Dans cet article, nous considérons "A" (116 individus décrits à 13 locus soit 32 modalités) comme un exemple moyen de groupe ne présentant pas de structuration génétique évidente (c'est à dire pouvant être issu d'une seule et même unité panmictique recouvrant dans ce cas l'ensemble de l'aire de l'échantillonnage) alors que "B" (66 individus décrits à 13 locus soit 35 modalités) serait l'archétype d'un groupe structuré déjà hétérogène au niveau de la maille d'échantillonnage choisi. La figure 2 représente la projection des points-lignes (individus) d'une autre espèce de rongeurs analyése par A.F.C. sur un groupe de 81 individus décrits à 14 locus soit 25 modalités (C. MONGELARD, 1985). A la projection de ce groupe, que nous appelerons "C", nous avons superposé les enveloppes convexes représentant les deux populations étudiées (C1 et C2). Nous considérons ce groupe comme génétiquement structuré de part ses particularités chromosomiques. Dans cet échantillon "C", nous savons *a priori* que le tirage n'a pas été réalisé dans une seule et même urne, puisque "C1" et "C2" ont des formules chromosomiques différentes.

3. Résultats

3.1. Essai de comparaison de tableaux d'homologies alléliques

Partant d'un tableau Individus X Génotypes, l'idée consiste à calculer l'homologie allélique cumulée pour chaque paire d'individus. Il s'agit du nombre d'allèles identiques rencontrés chez deux individus (contrairement à l'analyse factorielle des correspondances cet indice n'introduit aucune pondération liée aux fréquences différentes des allèles).

	Locus 1	locus 2	locus 3	locus 4	locus 5	...	Locus N
Individu i	090100	100100	080100	100110	100100	...	090090
Individu j	100100	100110	080100	100100	080100	...	090090
Homologie	01	10	11	10	01	...	11
Cumul	01	22	34	55	56		78

La valeur du cumul des homologies alléliques fournit un indice (ou degré de ressemblance pour chaque paire d'individus comparés. Sur un tableau contenant X individus, nous construisons la loi de distribution des (X*(X-1))/2) indices d'homologie et en calculons la moyenne et la variance. A partir de 100 permutons, nous établissons la loi de distribution de la moyenne et de la variance du degré

d'homologie allélique. La moyenne et la variance du degré d'homologie de l'échantillon réel seront respectivement placées sur ces différentes distributions ce qui nous permettra de situer cet échantillon parmi ses homologues stochastiques.

Voici les résultats obtenus sur, d'une part, l'échantillon A supposé non structuré et l'échantillon B supposé structuré, et d'autre part sur l'échantillons structuré C.

	GROUPE A		GROUPE B		GROUPE C	
	Homologie moyenne	Variance	Homologie moyenne	Variance	Homologie moyenne	Variance
Minimum observé	11 30	1 40	11 41	1 32	11 50	1 39
Maximun observé	11 64	1 89	11 70	2 14	11.90	1 55
Valeur du cas réel	**11.16**	**2.21**	**11.27**	**2.70**	**11.42**	**1.82**

Dans les trois échantillons, la valeur de l'homologie moyenne du cas réel reste inférieure à celles qui ont pu être calculées sur les permutons, les individus réels (en moyenne) se ressemblent moins entre eux que les individus d'un permuton. Ce test ne permet cependant pas de différencier les trois groupes.

3.2. Essai de comparaison de tableaux par identité génétique

Définition :

. Soit une population de N individus, soit a_{ij} le $i^{ème}$ allèle (i=1,..., A) au $j^{ème}$ locus (j=1,..., L).

. Soit dans cette population deux individus X et Y . Soit P_{ijk}, la probabilité de tirer l'allèle i au locus j chez l'individu k.

. Soit F_{aij} la fréquence de cet allèle dans l'échantillon.

Nous définissons entre deux individus différents X et Y l'identité génétique I(X,Y) :

$$I(X,Y) = \left[\sum_j \left[\sum_i P_{ijx}.P_{ijy} / \sum_i F_{aij}^{\,2}\right]\right] / \sum_j \frac{1}{\sum_i F_{aij}^{\,2}}$$

qui représente la probabilité moyenne de tirer deux allèles identiques chez l'un et l'autre individu, probabilité pondérée par l'homozygotie rencontrée à chaque locus.

	GROUPE A		GROUPE B		GROUPE C	
	Identité moyenne	Variance	Identité moyenne	Variance	Identité moyenne	Variance
Minimum observé	0.6800	0.0045	0 6666	0 0043	0 7863	0.0036
Maximun observé	0 6803	0 0056	0 6671	0 0067	0 7866	0 0042
Valeur du cas réel	**0.6798**	**0.0107**	**0.6655**	**0.0141**	**0.7861**	**0.0058**

Comme précédemment, l'identité génétique moyenne du cas réel reste inférieure à celles calculées sur les permutons pour les trois échantillons et la variance demeure plus importante pour les collections réelles.

3.3. Essais de comparaison de tableaux par étude des probabilités individuelles

Définition :

. Soit une population de N individus, soit a_{ij} le $i^{ème}$ allèle (i=1 à A) au $j^{ème}$ locus (j=1 à L).

6. Soit F_{aij} la fréquence de cet allèle dans l'échantillon.

. Soit dans cette population un individu X .

La probabilité de tirage d'un individu X issu de l'urne déduite, sous l'hypothèse de panmixie et d'indépendance des caractères, s'énonce ainsi :

$$P(X) = \pi_j Faij\bullet$$

• : $F_{aij}{}^2$ si le locus j est à l'état homozygote, $2 * F_{aij} * F_{ai'j}$ si le locus j est à l'état hétérozygote pour les allèles i et i'.

	GROUPE A		GROUPE B		GROUPE C	
	Probabilité moyenne	Variance	Probabilité moyenne	Variance	Probabilité moyenne	Variance
Minimum observé	0.81 10^{-3}	0 89 10^{-6}	0 36 10^{-3}	0 48 10^{-6}	8 33 10^{-3}	8.00 10^{-5}
Maximun observé	1.21 10^{-3}	2 57 10^{-6}	0 86 10^{-3}	2 35 10^{-6}	1 20 10^{-2}	1.46 10^{-4}
Valeur du cas réel	**0.83 10^{-3}**	**1.12 10^{-6}**	**0.35 10^{-3}**	**0.30 10^{-6}**	**8.53 10^{-3}**	**9.70 10^{-5}**

Considérons les probabilités moyennes individuelles dans un échantillon : un échantillon dit structuré aurait une probabilité moyenne plus faible que celle de ses homologues stochastiques, c'est le cas du groupe B. A l'inverse un échantillon peu structuré, donc plus proche d'un sous ensemble d'idéal panmictique, verrait

sa probabilité moyenne incluse dans l'intervalle de confiance des probabilités de ses homologues stochastiques, c'est le cas du groupe A qui apparaît à 7 % de la valeur minimale des permutons, c'est-à-dire que 93 des 100 permutons lui sont supérieurs. Quant au groupe C, il apparaît comme un peu plus structuré que A, à 1 % de sa valeur minimale.

3.4. Contre épreuve

Il apparaît que nous disposons maintenant de trois indices bien plus sensibles que ce qui est proposé par la méthode STATIS avec l'utilisation des coefficients RV par exemple, peut-être même trop sensibles puisqu'ils rejettent l'hypothèse nulle dans tous les cas, sauf le dernier où nos échantillons sont classés dans l'ordre B, C, A. A apparaît comme le moins structuré puisque le dernier test le place au dessus du seuil de significativité, dans le $7^{ième}$ percentile de la distribution et C est lui voisin de la borne des 1 % . On peut donc se poser légitimement la question de savoir si l'on peut sur ces bases définir un test qui ne rejette pas trop facilement l'hypothèse nulle. La contre-réponse consiste à montrer pour un test donné qu'il existe bien des situations où des échantillons naturels se trouvent près du centre de gravité de la distribution. Avec l'échantillon C, nous avons la chance de disposer d'un cas où les deux sous échantillons C1 et C2 (suffisamment nombreux) correspondent à des populations bien précises. Nous leur avons donc appliqué séparement les tests basés sur les trois indices et obtenus les résultats suivants :

	GROUPE C1		GROUPE C2	
	Homologie moyenne	Variance	Homologie moyenne	Variance
Minimum observé	11 57	1.03	11 83	0.97
Maximun observé	12 20	1.61	12 19	1 63
Valeur du cas réel	**11.72**	**1.42**	**11.92**	**1.29**

	GROUPE C1		GROUPE C2	
	Identité moyenne	Variance	Identité moyenne	Variance
Minimum observé	0.7772	0.0026	0 8284	0.0027
Maximun observé	0 7785	0.005 1	0 8290	0.0043
Valeur du cas réel	**0.7775**	**0.0043**	**0.8285**	**0.0039**

	GROUPE C1		GROUPE C2	
	Probabilité moyenne	**Variance**	**Probabilité moyenne**	**Variance**
Minimum observé	7.60 10 -3	5.96 10 -5	1.84 10 -2	4.06 10 -4
Maximun observé	1.19 10 -2	1.11 10 -4	2.89 10 -2	7.57 10 -4
Valeur du cas réel	**1.00 10 -2**	**9.59 10 -5**	**2.19 10 -2**	**5.69 10 -4**

Il apparaît sur ces tableaux que C1 et C2 sont nettement plus homogènes que leur somme. La valeur de l'homologie moyenne de chacun des cas réels se situe mainteant dans l'intervalle délimité par la valeur minimale et la valeur maximale. Nous pouvons même calculer que le cas réel C1 est à 14 % de sa valeur minimale, c'est-à-dire que 86 des 100 permutons lui sont supérieurs et que le cas réel C2 est à 8 % de sa valeur minimale. Pour la distance génétique la situation est la même que précédemment dans la mesure où le cas réel C1 est à 8 % de sa valeur minimale est que le cas réel C2 est à 3 % de sa valeur minimale. Quant à l'indice probabiliste, les cas réels sont à environ 50 % de leur valeur minimale; ces résultats sont donc en conformité avec l'idée que les généticiens se font de l'absence de structuration.

Discussion et conclusion

Le recours à la simulation d'échantillons aléatoires a semblé utile à la caractérisation des structures spatio-temporelles de l'espèce grâce aux correspondances partielles entre les allèles des gènes polymorphes. Nous avons essayé d'estimer les distorsions affectant les données brutes par rapport à leurs homologues stochastiques.

L'hypothèse réaliste de panmixie nous a amenés à engendrer des corpus de données aléatoires les plus proches possibles, d'un point de vue informationnel, de nos échantillons réels; nous avons préféré le "permuton" au "simulon" car ce dernier ne représentait que partiellement l'homologue stochastique de la collection réelle. Ce choix étant réalisé, nous avons testé plusieurs méthodes rétrodiagnostiques qui ont présenté divers niveaux de sensibilité. Nous avons, dans ces essais, décidé de nous munir de trois groupes d'individus : l'un supposé non structuré (groupe A) et deux autres supposés structurés (groupes B et C). On peut alors se demander, sachant déjà définir a priori cette notion de structuration, s'il est bien nécessaire de chercher à mettre au point des tests compliqués de validation. La réponse est positive, car nos critères de structuration étaient, dans ce cas précis, confortés par quelques observations d'ordre écologique ou cytologique nous permettant de considérer le groupe B comme un ensemble hétérogène d'unités à répartition géographique morcelée et le groupe C comme constitué de deux unités panmictiques (C1 et C2). Le groupe A quant à lui montrait une homogénéité certaine. Il est malheureusement assez rare dans notre domaine de disposer d'informations

complémentaires permettant de supposer a priori de l'hétérogénéité d'un groupe; il est donc important de se doter d'outils adaptés de validation statistique.

Nous avons utilisé différents indices (homologie allélique, identité génétique et probabilité) à travers la comparaison de lois statistiques; seule l'approche probabiliste semble donner les résultats conformes à l'idée que les généticiens se font de la structuation et se trouve être rapide d'utilisation. Notre hypothèse nulle H_O étant : l'échantillon observé est homogène et donc supposé tiré d'une unité panmictique, ceci nous amène à discuter de la sensibilité des tests précedemment employés :

Le test sur les indices d'homologie ainsi que le test basé sur l'identité génétique nous poussent à rejeter l'hypothèse H_O quel que soit l'échantillon A, B ou C considéré et donnent de très mauvais scores pour les sous-échantillons C1 et C2. Ces deux tests semblent donc trop sensibles dans la mesure où il nous font rejeter l'hypothèse nulle pour l'échantillon A alors que cette dernière est dans ce cas vraisemblable.

Le dernier test enfin, utilisant un indice probabiliste, permet de rejeter l'hypothèse nulle dans le cas du groupe B, de l'accepter de justesse dans le cas du groupe A et de pouvoir la rejeter au seul de 1 % pour le groupe C, alors que la contre-épreuve réalisée sur C1 et C2 désigne ceux-ci comme homogènes.

Dans l'élaboration de tests, le problème délicat est de savoir avec quel degré de sensibilité veut-on comparer les échantillons réels avec les prédictions attendues sous l'hypothèse nulle. Il n'y a pas *a priori* de réponse unique préétablie à cette question. Pour prendre une analogie physique, si l'on s'intéresse à l'écart de pièces sorties d'usine par rapport à un modèle circulaire idéal, il est clair que l'on pourra toujours prouver l'existence de petites déformations si l'on peut raffiner indéfinimment la précision de l'outil de mesure. Dans le cas qui nous préoccupe, ceci est clairement mis en évidence : parmi les approches employées, certaines se sont avérées beaucoup trop peu sensibles (par exemple la méthode STATIS, c.f. § 2.2) alors que d'autres (indices des § 3.1 et 3.2) révèlent que les trois échantillons naturels testés montrent des petites singularités par rapport à un idéal panmictique. Seul le troisième indice (§ 3.3) tend à ramener un échantillon réputé non structuré, comme A par exemple, dans la fourchette des cas simulés, assez loin du centre de la distribution cependant, et subit une contre-épreuve avec succès dans le cas de C1 et C2.

A ce stade, il convient cependant de s'interroger sur les significations de ces écarts et de se demander jusqu'à quel point nous devons en tenir compte. Cette discussion nécessite de revenir à la réalité biologique de l'échantillonnage : il existe de nombreux facteurs qui peuvent modifier, ne serait-ce que de manière infime, la structure de cet échantillonnage par rapport à un idéal panmictique. Dans le cas des espèces que nous avons considéré ici, nous pouvons penser en premier lieu à l'*apparentement possible* des individus analysés : dans la nature, les accouplements ne se font pas au hasard et au cours de l'échantillonnage, des individus apparentés peuvent avoir été capturés. Dans un petit échantillon, ceci peut être suffisant pour créer des déséquilibres d'associations qui seront détectables par nos coefficients, coefficients sûrement assez sensibles au fait que deux modalités se retrouvent par exemple simultanément présentes dans deux individus apparentés

plus que ne le laisserait prévoir une stricte indépendance. Ceci étant dit, l'arbre ne doit pas cacher la forêt, et nous pensons pouvoir justifier les conclusions suivantes : les structurations qu'il nous intéresse de mettre en évidence sont *a priori* du type de celles présentées dans l'échantillon B ou l'échantillon C. L'indice probabiliste est donc celui qui discrimine le mieux entre la situation B et celle des autres échantillons analysés. C'est donc celui que nous retiendrons et développerons dans des améliorations futures. Le risque de seconde espèce (rejet de H_1 alors qu'elle est vraie) est quasiment nul du fait de la sensibilité de ce coefficient, tandis que le risque de première espèce (rejet de H_O alors qu'elle est vraie) est laissé à l'appréciation de l'expérimentateur en fonction de la position de l'échantillon réel par rapport à la distribution des permutons, position qui est assortie d'une probabilité d'occurence sous l'hypothèse nulle. Nous pensons également que cette approche peut être généralisable à d'autres types de données dès lors que l'hypothèse nulle de non-association des variables entre elles est réaliste.

Bibliographie

AUTEM, M. et coll., 1987 - Entre l'individu et l'espèce toute entière, les panmixons existent-ils ? Les problèmes de l'analyse de l'organisation de l'espèce. *Actes du Coll.Nat. CNRS "Biologie des populations" Lyon : 4-6 septembre 1986.*

BENZECRI, J.P. et coll., 1973 - Analyse des données, *Tome 1 et 2, Dunod.*

DOBZHANSKY, Th., 1951 - Genetics and the origin of species, *Third Edition, New York Columbia University Press, p15.*

DUPLANTIER J.M., 1988 Thèse dr. d'état - Biologie Evolutive de populations du genre Mastomys (Rongeur, Muride) au Sénégal, *Université des Sciences et Techniques du Languedoc, Montpellier.*

EFRON, B., 1979 - Bootstrap methods : An other look at the Jackknife, *Annals of statistics, 7, 1-26.*

EFRON, B., 1982 - The jackknife, the bootstrap and other resampling plans, *Society for Industrial & Appl. Mathematics. Philadelphia, PA.*

ESCOUFIER, Y., 1973 - Le traitement des variables vectorielles, *Biometrics, 29, pp 751-760.*

GAUCH, H.G., jr, 1982 - Multavariate Analysis in Community Ecology, *Cambridge.U.Press, p. 1952.*

GREENACRE, M.J., 1983 - Theory and applications of correspondance Analysis, *Academic Press, New York.*

HERNANDEZ, J.L., and WEIR, B.S., 1989 - A disequilibrium coefficient approach to Hardy-Weinberg testing . *Biometrics 45, 53-70.*

HOLMES JUNCA, S., 1985 - Thèse 3ème cycle - Outils informatiques pour l'évaluation de la pertinence d'un résultat en Analyse des Données. *Université des Sciences et Techniques du Languedoc, Montpellier.*

LAVIT, C., 1988 - Analyse conjointe de plusieurs tableaux de données, *Masson.*

LEARMONTH, G.P. and LEWIS, P.A.W., 1973 - Naval Postgraduate School Random Number Generator Package LLRANDOM, NPS55LW7306LA, *Naval Postgraduate School, Monterey, California.*

LEHMER, D.H., 1951 - Mathematical methods in Large-scale units, *Ann. Comp. Lab., Havard University, 26, 141-146.*

MILLER, R.G., 1974 - The Jackknife - A review, *Biometrika, Vol.61, 1, 1-15.*

MONGELARD, C., 1985 - Thèse 3ème cycle - Structures géniques et chromosomiques des populations de souris robertsonnienes de Mus musculus domesticus en Italie du Nord. Discussion du modèle de spéciation stasipatrique, *Université des Sciences et Techniques du Languedoc, Montpellier.*

ROUX, M. et LEBRETON, J.M., 1988 - Progiciel BIOMEO, *Groupe de Biométrie, C.E.P.E. C.N.R.S., Montpellier.*

SHE, X. et coll., 1987 - Multivariate analysis of genetic exchanges between Solea aegyptiaca and Solea senegalensis (Teleosts, Soleidae), *Biol. Jour. of the linnean Society, Vol. 32 : 357-371.*

STONE, M., 1974 - Cross-Validatory Choice and Assessment of Statistical Preductions (with discussion), *J.Roy.Stat.Soc.B. Vol.36, 111-148.*

TUCKEY, J.W. and MALLOWS, C., 1982 - An Overview of Techniques of Data Analysis Emphasizing its Exploratory Aspects, *in Somme Recent Advances in Statistics, ed Epstein.*

TABLEAU Ia
Matrice de données brutes

Nom	Locus A	Locus B	Locus C	Locus D	Locus E	Locus F	
NY01	080080	100100	100100	140140	100100	100100	...
NY03	080080	100100	100100	100100	100100	100100	...
NY04	080100	100100	100100	100140	100100	100120	...
NY05	080100	100100	100100	100140	100100	100100	...
NY06	080080	100100	100100	100120	100100	100100	...
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	
.	.	.	.	.	.	.	

TABLEAU Ib
Matrice de données codées

Nom	Locus A	Locus B	Locus C	Locus D	Locus E	
NY01	[2 0 0]	[0 2]	[2 0]	[0 0 2]	[0 0 2]	...
NY03	[2 0 0]	[0 2]	[2 0]	[2 0 0]	[0 0 2]	...
NY04	[1 1 0]	[0 2]	[2 0]	[1 0 1]	[0 0 2]	...
NY05	[1 1 0]	[0 2]	[2 0]	[1 0 1]	[0 0 2]	...
NY06	[2 0 0]	[0 2]	[2 0]	[1 1 0]	[0 0 2]	...
. . .	. .	. .	. .	. . .	. . .	
. . .	. .	. .	. .	. . .	. . .	
. . .	. .	. .	. .	. . .	. . .	

Légende: l'individu NY04 posséde pour le locus D (caractère) les deux allèles (modalités) respectivement symbolisés 100 et 140 dans le tableau Ia, représenté par [1 0 1] dans le tableau Ib.

FIGURE 1a. – *A.F.C. Groupe A*

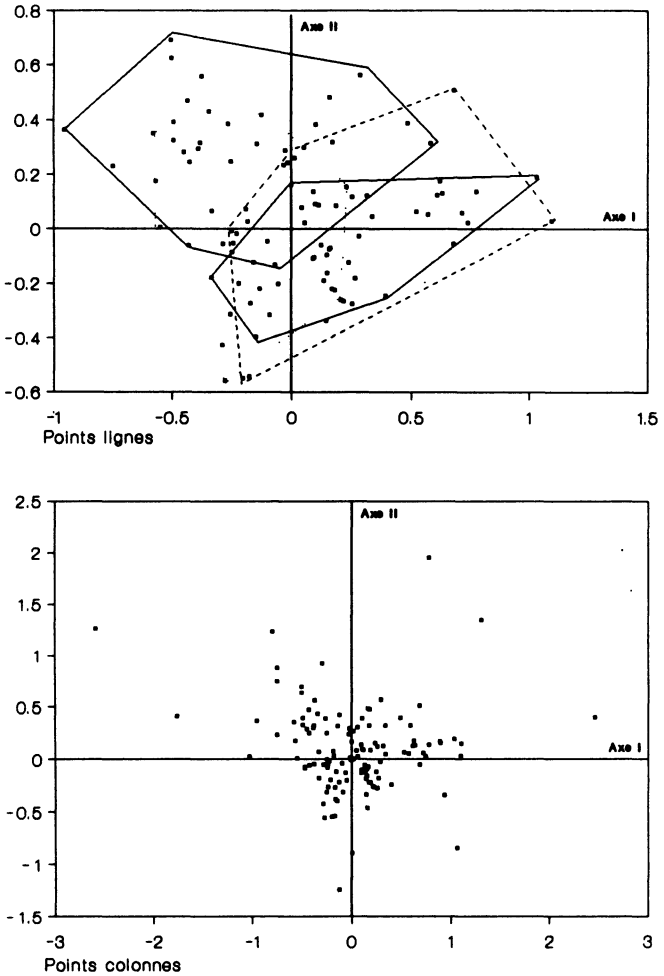
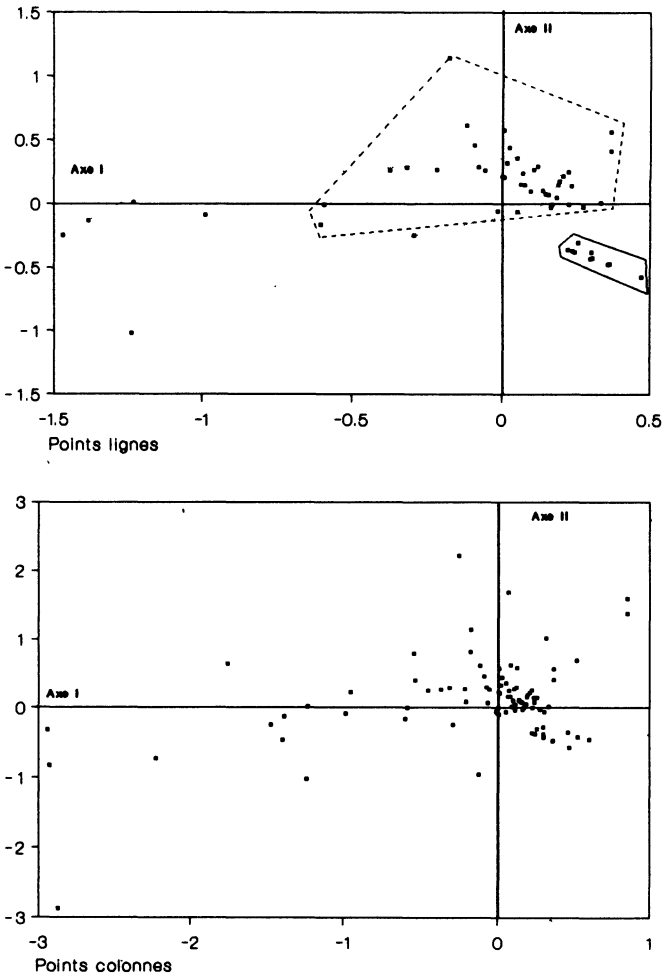


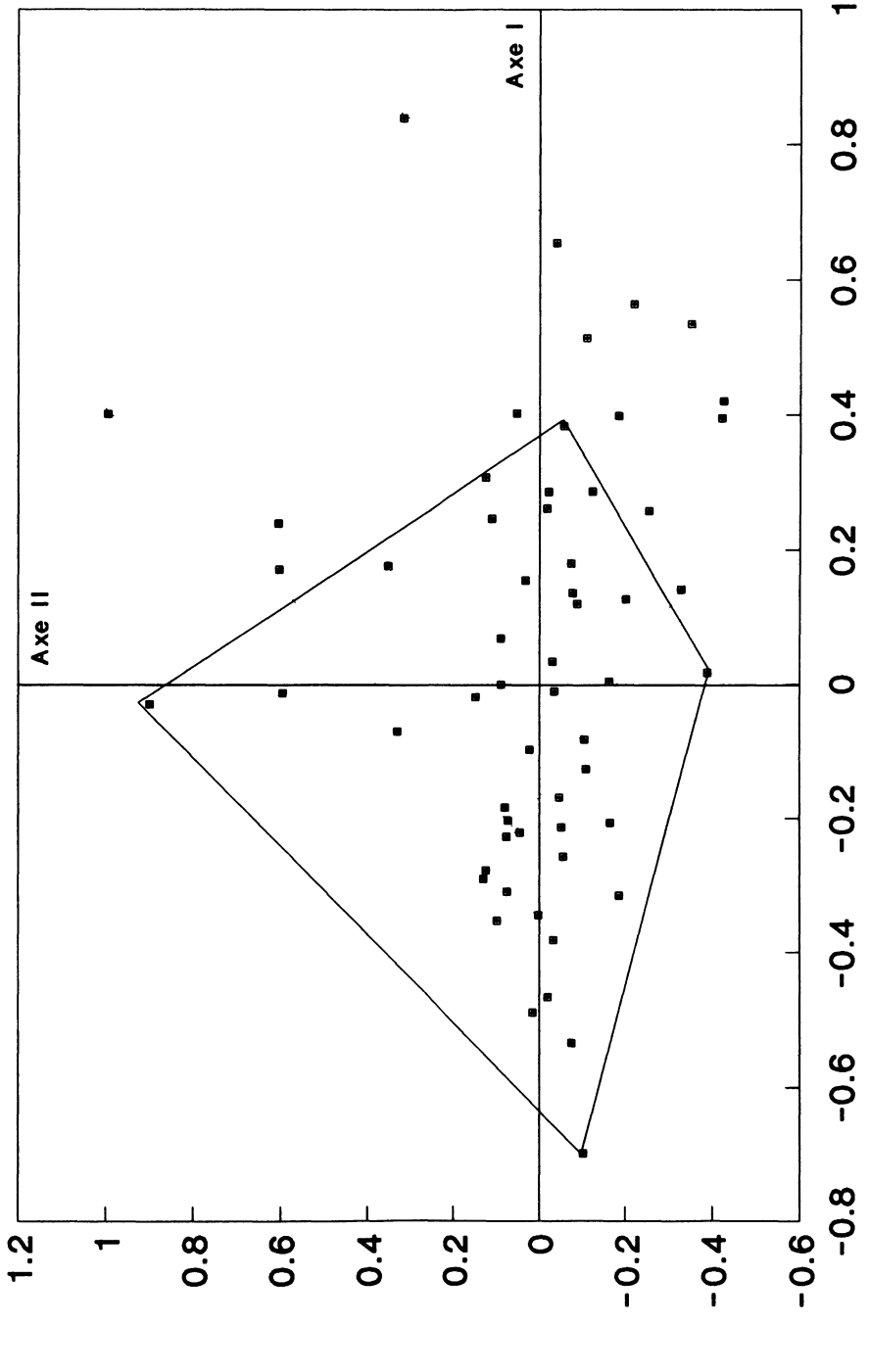
FIGURE 1b. – *A.F.C. Groupe B*



Deux cas typiques d'analyse des correspondances d'échantillons réels

Légende : *La figure 1a représente un échantillon non structuré malgré des provenances géographiques différentes ; la figure 1 représentant une structure particulière le long de l'axe 1.*

FIGURE 2. – *A.F.C. Groupe C*



Analyse factorielle des correspondances d'un échantillon réel

Légende : *Cette figure représente un échantillon structuré constitué de deux sous échantillons à formules chromosomiqu différentes.*

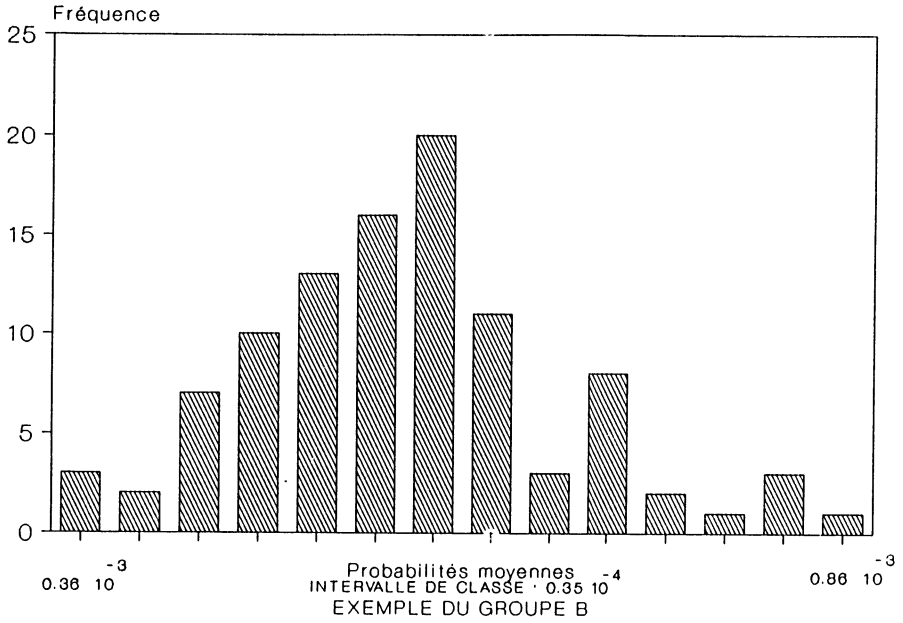


FIGURE 3
Loi de distribution des probabilités moyennes sur 100 tableaux permutés