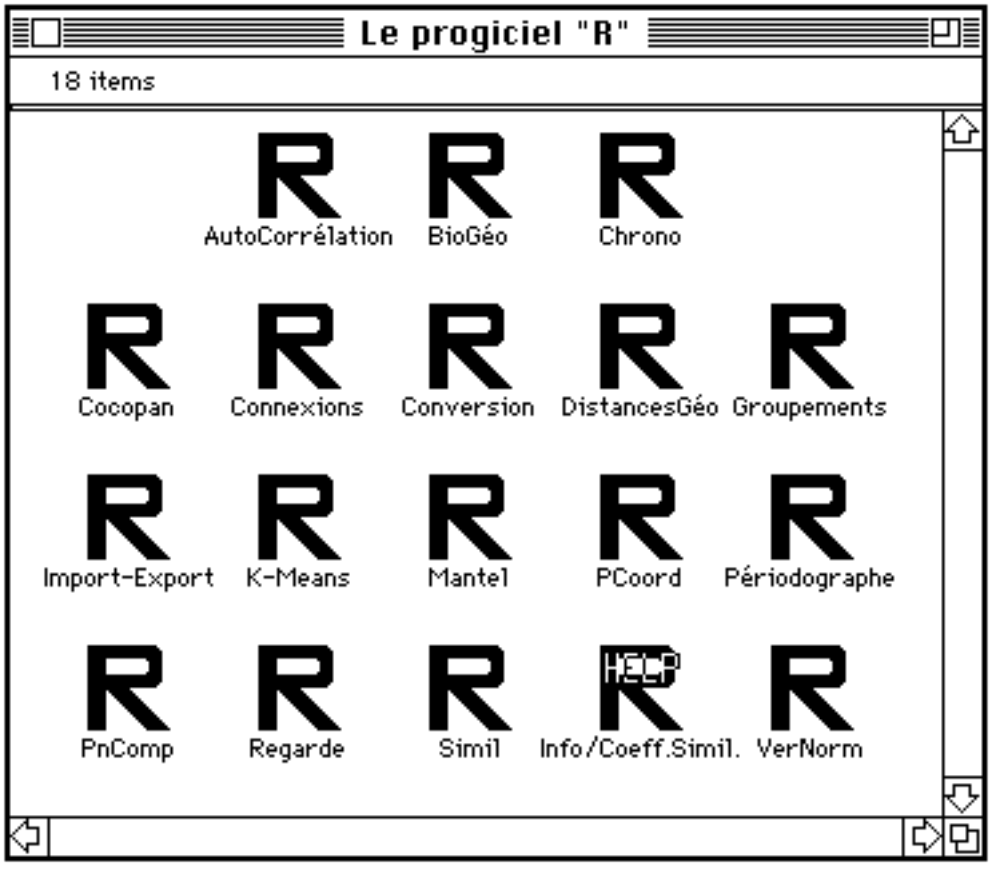
Le progiciel R

Analyse multidimensionnelle, analyse spatiale

Versions CMS (IBM), VMS (VAX) et Macintosh

Pierre Legendre / Alain Vaudor



Université de Montréal

Le progiciel R

Analyse multidimensionnelle, analyse spatiale

Versions CMS (IBM), VMS (VAX) et Macintosh

Pierre Legendre et **Alain Vaudor**

Département de sciences biologiques
Université de Montréal
C.P. 6128, Succursale A
Montréal, Québec
Canada H3C 3J7

Courrier électronique — P. Legendre: Legendre @ Ere.UMontreal.CA
A. Vaudor: Vaudor @ Ere.UMontreal.CA

Ce manuel a été préparé avec l'assistance éditoriale de
Chantal Ouimet, **François-Joseph Lapointe** et **Gilles Lavoie**

Université de Montréal, septembre 1991
Mise à jour:

Avertissement

Ces programmes vous sont fournis sans aucune garantie implicite ou explicite de bon fonctionnement. Il s'agit de programmes mis au point dans le cadre de recherches universitaires. Cependant, si vous éprouvez des problèmes avec l'un ou l'autre des programmes de ce progiciel, nous serons heureux de tenter de vous dépanner (voir section 5, page 6). Les chercheurs peuvent utiliser ces programmes pour les fins de leurs recherches, mais le code-source des programmes demeure la propriété des auteurs de ce manuel.

Vous devez disposer des polices de caractères suivantes pour imprimer ce document: *Times*, *Courier* et *Symbol*. La mise en page a été effectuée en fonction d'une imprimante laser de type PostScript.

Référence de ce manuel:

Legendre, P. et A. Vaudor. 1991. Le progiciel R — Analyse multidimensionnelle, analyse spatiale. Département de sciences biologiques, Université de Montréal. iv + 144 p.

Table des Matières

UN PEU D'HISTOIRE

Cet ensemble de programmes d'ordinateur a été écrit au fil des ans par Alain Vaudor (Analyste de l'Informatique) et Pierre Legendre. Le développement du progiciel débuta en 1978, à l'Université du Québec à Montréal, sur machines PDP-10 et CDC/CYBER. En 1980, le progiciel déménagea en même temps que nous à l'Université de Montréal, où son développement s'est poursuivi depuis. Des programmes furent d'abord mis au point pour les méthodes générales d'analyse de données (mesures de similarité et de distance, différentes méthodes de groupement, des ordinations, etc., en plus des programmes utilitaires nécessaires); les programmes correspondant à des méthodes plus spécifiques, répondant à des questions plus particulières, furent développés ensuite (périodogramme de contingence, groupement chronologique, groupements avec contrainte de contiguïté spatiale, autocorrélation spatiale, tests de Mantel, Cocopan). Les programmes furent graduellement améliorés et devinrent plus conviviaux, grâce aux commentaires de générations successives d'étudiants diplômés et d'autres usagers. De premières versions pour machines IBM furent mises au point, indépendamment, à l'University of Waterloo (Ontario) et à l'Université de Sherbrooke (Québec), pour utilisation en lot seulement. La version conversationnelle IBM a été développée par P. Legendre depuis 1985, d'abord sur les ordinateurs du C.N.U.S.C. (Montpellier, France) et du Department of Ecology and Evolution, State University of New York (Stony Brook, U.S.A.), puis sur celui de l'École Polytechnique de Montréal. Cette version fut adaptée au VAX à l'Université de Montréal en 1989. Les programmes devinrent bilingues (français/anglais) à l'occasion de l'implantation à Stony Brook. Il aura fallu 13 ans pour compléter le développement de ce progiciel et la rédaction de sa documentation; cette période inclut le temps nécessaire au développement, en notre laboratoire, de plusieurs des méthodes qui y sont mises en oeuvre, ainsi que la rédaction des publication concomitantes.

Les programmes eux-mêmes sont écrits en PASCAL alors que les programmes d'appel sur IBM sont en REXX et en DCL sur VAX. Ils ont été fournis à nombre d'établissements universitaires en Amérique du Nord, en Europe et en Amérique du Sud. Les versions disponibles en ce moment sont:

Type d'ordinateur	Conversation avec l'usager	Système d'exploitation	Programmes d'appel
IBM (grands ordinateurs)	Français ou anglais	VM/CMS	Fichiers EXEC (REXX)
VAX	Français ou anglais	VAX/VMS	Fichiers DCL
Apple Macintosh	Français ou anglais		Cliquez sur l'icône!

On peut se procurer ces programmes contre 25 $ (Can., US ou Aust.), ce qui couvre le prix de la disquette et d'une copie de la documentation ainsi que les frais de poste. Précisez la version désirée; pour les versions CMS et VMS, indiquez si vous désirez recevoir une disquette devant être relue par un Macintosh ou par un micro-ordinateur opérant sous MS/DOS (si vous préférez des disquettes de 5.25 pouces, précisez-le). Une copie de la documentation accompagnera tout envoi; spécifiez la langue désirée (français ou anglais). Des programmes individuels pourront être expédiés par courrier électronique. La version Macintosh est fournie déjà compilée, alors que les versions pour grands ordinateurs sont fournies sous la forme de fichiers-source, ce qui permet aux usagers de changer la taille des matrices pouvant être traitées par les programmes, ainsi que la langue de la conversation; ceci implique cependant que les usagers doivent compiler eux-mêmes les programmes avant de pouvoir les utiliser (compilateur PASCALVS ou VSPASCAL sur IBM; compilateur PASCAL sur VAX).

Le nom du progiciel, "R", provient de nos travaux sur machine PDP-10 en 1978. Sur ce type de machine, "R" (pour *Run*) est la commande de démarrage d'un programme. Sur les machines Control Data, "R" était un grand fichier de commande en langage CCL, à partir duquel l'usager pouvait mettre en marche n'importe quel programme du progiciel; cette façon de faire simulait la façon de procéder sur machine PDP. Le nom de ce fichier s'est imposé pour devenir le nom du progiciel.

ACCES AUX PROGRAMMES

1. En interactif, système CMS (IBM)

Pour utiliser les programmes de ce progiciel à partir de sa propre machine virtuelle, l'usager doit d'abord s'attacher au minidisque contenant les fichiers EXEC et les programmes constituant "R", à moins qu'il ne travaille directement sur la machine virtuelle contenant tous ces fichiers.

Inscrire ici les commandes nécessaires sur votre machine:

Les commandes EXEC disponibles sont les suivantes. Chacune provoque l'exécution du programme correspondant.

- **AUTOCOR**
- **BIOGEO**
- **CHRONO**
- **COCOPAN**
- **CONVERT**
- **DIST**
- **EXPNTS**
- **EXPORT**
- **IMPORT**
- **INTERLNK**
- **KMEANS**
- **LANCE**
- **MANTEL**
- **PCOORD**
- **PERIOD**
- **REGARDE**
- **SIMIL**
- **VERNORM**

Ces commandes mettent en route les programmes suivants:

- **AUTOCOR**: Autocorrélation spatiale unidimensionnelle (coefficients *I* de Moran et *c* de Geary). Ce programme permet également de calculer une liste de liens selon différents algorithmes, utilisée par les programmes Biogeo, KMeans (lorsqu'il est employé avec contrainte) et Cocopan.
- **BIOGEO**: Groupement avec contrainte de contiguïté spatiale. Méthode: liens intermédiaires.
- **CHRONO**: Groupement chronologique (avec contrainte de contiguïté temporelle, ou spatiale en une seule dimension).
- **COCOPAN**: Analyse de variance en présence d'autocorrélation spatiale.
- **CONVERT**: Convertit les **S**imilarités en **D**istances, ou les **D**istances en **S**imilarités.
- **DIST**: Calcul des distances en suivant la courbure de la terre, à partir de longitudes et de latitudes.
- **EXPNTS**: Convertit une matrice binaire de type SIMIL en une matrice binaire de type NT-SYS (Numerical Taxonomy and Multivariate Analysis System de F. James Rohlf).
- **EXPORT**: Convertit une matrice binaire de type SIMIL en une matrice ASCII carrée.
- **IMPORT**: Convertit une matrice ASCII carrée en une matrice binaire de type SIMIL.
- **INTERLNK**: Groupement à liens intermédiaires (algorithme de liaison proportionnelle).
- **K-MEANS**: Groupement selon la méthode K-Means (variance minimum), avec ou sans contrainte de contiguïté spatiale.
- **LANCE**: Groupement selon l'algorithme général de Lance & Williams, incluant Ward.
- **MANTEL**: test de Mantel, tests partiels de Mantel, corrélogramme multidimensionnel.
- **PCOORD**: Analyse en coordonnées principales.
- **PERIOD**: Calcul du périodogramme de contingence.
- **REGARDE**: pour regarder ou imprimer un fichier binaire produit par SIMIL.
- **SIMIL**: 50 mesures de ressemblance. Les coefficients sont calculées uniquement entre les LIGNES d'un fichier de données. Pour les coefficients en mode Q, les lignes de la matrice de données doivent correspondre aux objets; en mode R, les lignes doivent correspondre aux descripteurs.
- **VERNORM**: Vérification et normalisation des colonnes (variables) d'un fichier de données.

Certains programmes requièrent plus de mémoire que la quantité attribuée par défaut aux usagers. Ce problème peut aussi surgir si on a augmenté les dimensions d'un programme pour traiter des fichiers de données particulièrement grands. L'usager doit alors recourir a la commande DEF STOR pour avoir accès à de l'espace-mémoire supplémentaire.

Par ailleurs, lors de l'exécution des programmes conversationnels, le texte affiché par les programmes de même que les réponses de l'usager aux questions apparaissent normalement uniquement à l'écran. Enfin, les programmes CHRONO, MANTEL et PERIOD par exemple ne présentent qu'à l'écran le résultat de leurs calculs. Si on désire conserver cet ensemble de questions, de réponses et de résultats dans un fichier, en vue de le consulter ou de le faire éventuellement imprimer, il faut donner la commande suivante avant de démarrer l'exécution du programme:

```
CP SPOOL CONS START TO *
```

Cette commande doit être exécutée *en dehors de tout FILELIST*. De nouveau, on peut préférer inscrire à l'avance cette commande dans un fichier EXEC (appelé par exemple le fichier RETIENS EXEC). Après avoir fait exécuter un ou plusieurs programmes, et de nouveau *en dehors de tout FILELIST*, on écrit:

```
CLOSE CONS NAME MEMOIRE CONSOLE
CP SPOOL CONS STOP
```

(ces commandes peuvent se trouver dans un fichier EXEC). Le fichier contenant les interactions, auquel on donne par exemple le nom *MEMOIRE CONSOLE* comme ci-dessus, se retrouve dans le "Reader list", auquel on accède par la commande RDRL. On peut évidemment éditer ce fichier pour lui enlever des sections inutiles, avant de le faire imprimer.

2. En interactif, système VMS (VAX)

Sur machine VAX, les programmes sont appelés par des fichiers de commandes DCL équivalents aux EXEC de l'IBM: VERNORM.COM, SIMIL.COM, etc. L'usager possédant une copie du progiciel "R" sur son propre compte peut donc appeler directement les programmes en tapant le signe @ suivi du nom du programme désiré; par exemple: @VERNORM, @SIMIL, etc.

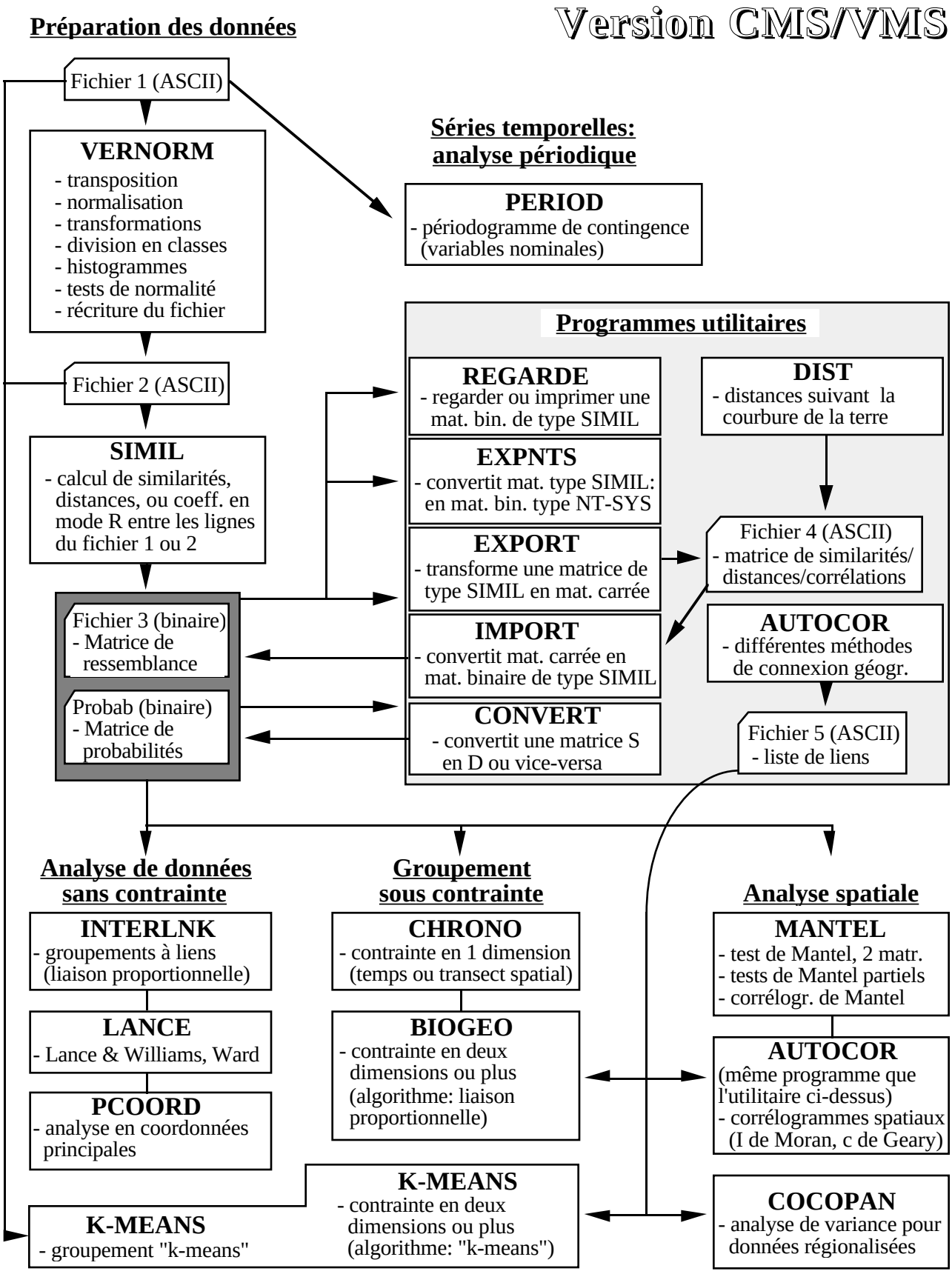
Une deuxième possibilité consiste à activer le fichier de commande R.COM en tapant @R. Ce fichier donne les noms et adresses des auteurs du progiciel et énumère les programmes disponibles dans le progiciel "R". Son exécution permet ensuite à l'usager d'appeler les programmes sans le symbole @; par exemple: VERNORM, SIMIL, etc.

On peut installer le progiciel de façon à ce qu'il soit accessible aux autres usagers du VAX. Le "dépositaire" du progiciel devra modifier tous les fichiers de commandes (y compris R.COM) en ajoutant son adresse-machine partout où un programme ou un autre fichier de commande est appelé (par RUN ou par @). Par exemple:

```
@VERNORM        peut devenir    @DUA1:[Tartempion]VERNORM
RUN SIMIL       peut devenir    @DUA1:[Tartempion]SIMIL
```

Il demandera à chaque usager d'ajouter dans son fichier LOGIN.COM une instruction du type:

```
$ R:=="@DUA1:[Tartempion]R.COM"
```


Version CMS/VMS
Préparation des données
Fichier 1 (ASCII)
VERNORM
- transposition
- normalisation
- transformations
- division en classes
- histogrammes
- tests de normalité
- récriture du fichier
Fichier 2 (ASCII)
SIMIL
- calcul de similarités, distances, ou coeff. en mode R entre les lignes du fichier 1 ou 2
Fichier 3 (binaire)
- Matrice de ressemblance
Probab (binaire)
- Matrice de probabilités
Séries temporelles: analyse périodique
PERIOD
- périodogramme de contingence (variables nominales)
Programmes utilitaires
REGARDE
- regarder ou imprimer une mat. bin. de type SIMIL
EXPNTS
- convertit mat. type SIMIL: en mat. bin. type NT-SYS
EXPORT
- transforme une matrice de type SIMIL en mat. carrée
IMPORT
- convertit mat. carrée en mat. binaire de type SIMIL
CONVERT
- convertit une matrice S en D ou vice-versa
DIST
- distances suivant la courbure de la terre
Fichier 4 (ASCII)
- matrice de similarités/ distances/corrélations
AUTOCOR
- différentes méthodes de connexion géogr.
Fichier 5 (ASCII)
- liste de liens
Analyse de données sans contrainte
INTERLNK
- groupements à liens (liaison proportionnelle)
LANCE
- Lance & Williams, Ward
PCOORD
- analyse en coordonnées principales
K-MEANS
- groupement "k-means"
Groupement sous contrainte
CHRONO
- contrainte en 1 dimension (temps ou transect spatial)
BIOGEO
- contrainte en deux dimensions ou plus (algorithme: liaison proportionnelle)
K-MEANS
- contrainte en deux dimensions ou plus (algorithme: "k-means")
Analyse spatiale
MANTEL
- test de Mantel, 2 matr.
- tests de Mantel partiels
- corrélogr. de Mantel
AUTOCOR
(même programme que l'utilitaire ci-dessus)
- corrélogrammes spatiaux (I de Moran, c de Geary)
COCOPAN
- analyse de variance pour données régionalisées

Après avoir validé son LOGIN de commandes, le nouvel usager n'aura plus qu'à taper

```
R
```

ce qui fera apparaître le message d'entrée. Dès lors, pour la session VAX en cours, chaque programme pourra être appelé simplement par son nom.

3. En lot (*"Batch"*), système CMS (IBM)

Pour l'exécution en lot, les noms des fichiers de données et de résultats sont spécifiés dans des fichiers EXEC. Les réponses aux questions posées par le programme, après le message informatif "EXECUTION BEGINS ...", doivent se trouver dans un fichier de réponses dont le nom sera placé dans le fichier EXEC.

Quatre programmes demandent parfois d'être accessibles en lot lorsqu'on désire traiter des fichiers de grande taille; il s'agit de SIMIL, MANTEL, AUTOCOR et PCOORD. Les fichiers EXEC correspondant (SIMILOT, MANTELOT, AUTOLOT, PCOORLOT) peuvent se trouver sur la machine PROGICIEL-R où réside le progiciel. Pour une exécution en lot, il est nécessaire de copier l'EXEC désiré de la machine PROGICIEL-R vers la vôtre et d'y apporter les adaptations nécessaires. Un programme se lance en lot par la commande habituelle; par exemple:

```
SUBMIT SIMILOT (CPU ...
```

Exemple: fichier SIMILOT EXEC — /* Ces lignes sont des commentaires */

```
/* Fichier de lancement du programme SIMIL en lot. */
GLOBAL TXTLIB VSPASCAL
FI OUTPUT PRINTER
/* Nom du fichier contenant les réponses aux questions: */
FI INPUT DISK reponses simil a
/* Nom du fichier de donnees: */
FI ENTREEC DISK fichier donnees a
/* Nom du fichier contenant la matrice de ressemblance calculee par SIMIL: */
FI SORTIE DISK fichier binaire a
/* Nom du fichier contenant les matrices de similarites partielles: */
FI PART DISK fichier partiel a
/* Nom du fichier contenant la matrice des probabilites, s'il y a lieu: */
FI PROBAB DISK fichier probab a
/* La ligne suivante lance la version redimensionnee du programme SIMIL: */
"LOAD SIMILOT (START"
/* Changer si nécessaire le nom de la machine d'où émane cette passe en LOT: */
"SENDFILE fichier binaire a TO PROGICIELR"
"SENDFILE fichier partiel a TO PROGICIELR"
"SENDFILE fichier probab a  TO PROGICIELR"
```

Les noms des différents fichiers doivent être adaptés à vos données. Le fichier de réponses ne doit contenir que les réponses aux questions posées par les programmes pour cette passe précise.

Exemple de fichier de réponses aux questions du programme SIMIL:

```
Un titre de votre choix.
380       [nombre de lignes ou de blocs de lignes]
109       [nombre de colonnes]
```

```
N       [il n'y a pas de noms d'objets en col. 1-10]
S01     [code désignant le coefficient désiré]
5       [l'information sera codée "1" à partir de la valeur 5]
```

Une façon simple d'obtenir la liste des questions est de lancer l'exécution de manière interactive sur un fichier bidon ou sur une partie du fichier réel.

4. Version Macintosh

Dans la version Macintosh, les programmes sont essentiellement les mêmes que dans les versions CMS et VMS. Dans quelques cas, des réarrangements ont été réalisés qui permettent de tirer meilleur partie de l'interface-usager du Macintosh. Les programmes disponibles sont les suivants:

* **AUTOCORRÉLATION**: Autocorrélation spatiale unidimensionnelle (coefficients I de Moran et c de Geary).
* **BIOGÉO**: Groupement avec contrainte de contiguïté spatiale. Méthode: liens intermédiaires.
* **CHRONO**: Groupement chronologique (avec contrainte de contiguïté temporelle, ou spatiale en une seule dimension).
* **COCOPAN**: Analyse de variance en présence d'autocorrélation spatiale.
* **CONNEXIONS**: Calcule une liste de liens selon différents algorithmes. Cette liste est utilisée par les programmes Biogéo, K-Means (employé avec contrainte), Cocopan et Autocorrélation.
* **CONVERSION**: Convertit les **S**imilarités en **D**istances, ou les **D**istances en **S**imilarités (équivalent de CONVERT des versions CMS et VMS).
* **DISTANCES GÉOGRAPHIQUES**: Calcul des distances en suivant la courbure de la terre, à partir de longitudes et de latitudes.
* **GROUPEMENTS**: Liens intermédiaires, Lance & Williams, Ward (remplace LANCE et INTERLNK des versions CMS et VMS).
* **IMPORT-EXPORT**: Pour importer des matrices de ressemblance et les transformer en format binaire de type *SIMIL*, ou pour exporter des matrices produites par *SIMIL* vers d'autres programmes. Remplace IMPORT et EXPORT des versions pour grands ordinateurs.
* **K-MEANS**: Groupement selon la méthode K-Means (variance minimum), avec ou sans contrainte de contiguïté spatiale.
* **MANTEL**: test de Mantel, tests partiels de Mantel, corrélogramme multidimensionnel.
* **PCOORD**: Analyse en coordonnées principales.
* **PÉRIODOGRAPHE**: Calcul du périodogramme de contingence.
* **PNCOMP**: Analyse en composantes principales.
* **REGARDE**: pour regarder ou imprimer un fichier binaire produit par SIMIL.
* **SIMIL**: 50 mesures de ressemblance. Les coefficients sont calculés uniquement entre les LIGNES d'un fichier de données. Pour les coefficients en mode Q, les lignes de la matrice de données doivent correspondre aux objets; en mode R, les lignes doivent correspondre aux descripteurs.
* **VERNORM**: Vérification et normalisation des colonnes (variables) d'un fichier de données.

Pour l'utilisation courante, il est préférable de transférer les programmes sur disque rigide, ou encore de travailler avec deux disquettes; assurez-vous que votre environnement de travail comprend un SYSTEM FILE, une icône correspondant à votre type d'imprimante, ainsi qu'un éditeur de programmation (voir la raison plus bas).

Si vous désirez utiliser l'imprimante (par exemple, pour obtenir les résultats des groupements), assurez-vous que la disquette où se trouve le système contient au moins de 30 à 50K d'espace libre, ce qui permettra au système de créer ses fichiers temporaires lors de l'impression.

Les fichiers de données doivent être des matrices rectangulaires de nombres entiers ou réels, du type "texte seulement" (code ASCII). On peut les extraire en "texte seulement" de chiffriers ou de programmes de traitement de texte, ou mieux encore, on peut les fabriquer à l'aide d'un éditeur de programmation, tel que celui fourni sur la disquette. Les fichiers de données transférés par MODEM à partir de grands ordinateurs sont habituellement de type ASCII.

Pour sélectionner le fichier d'entrée d'un programme, il suffit de cocher "OUVRIR" après avoir noirci le nom du fichier désiré. Ne sont présentés que les fichiers de la disquette qui sont d'un type approprié pour le programme en question: fichiers "texte seulement" pour l'entrée de VERNORM, SIMIL, PÉRIODOGRAPHE et IMPORT-EXPORT (selon l'option); fichiers binaires de type "SIMIL" pour IMPORT-EXPORT (selon l'option) et pour la plupart des autres programmes.

Fichiers de sortie: pour les fichiers de sortie de SIMIL, on change le nom proposé et on coche "ENREGISTRER". Pour les programmes d'analyse de données, les sorties se font normalement sur l'imprimante. Si on est d'accord, on coche la case "ENREGISTRER"; sinon, il suffit de changer le mot "imprimante" pour un nom de fichier de son choix et de cocher la case "ENREGISTRER". Ce fichier, de type ASCII, pourra être relu à l'aide d'un éditeur de programmation.

Lorsqu'on doit fournir des nombres en réponse aux questions du programme, il faut se rappeler que les programmes sont écrits en PASCAL; il faut donc écrire " 0.5 " et non pas " .5 " , par exemple. Il en est de même des fichiers de données. Cette recommandation est également valable pour les versions CMS et VMS. Dans la version Macintosh, les programmes numérotés 3 et plus sont libérés de cette contrainte et peuvent lire des données du type .2, -.57, +0.1, -0., 5E+2, +1.0e-8, etc.

5. Documentation d'un problème

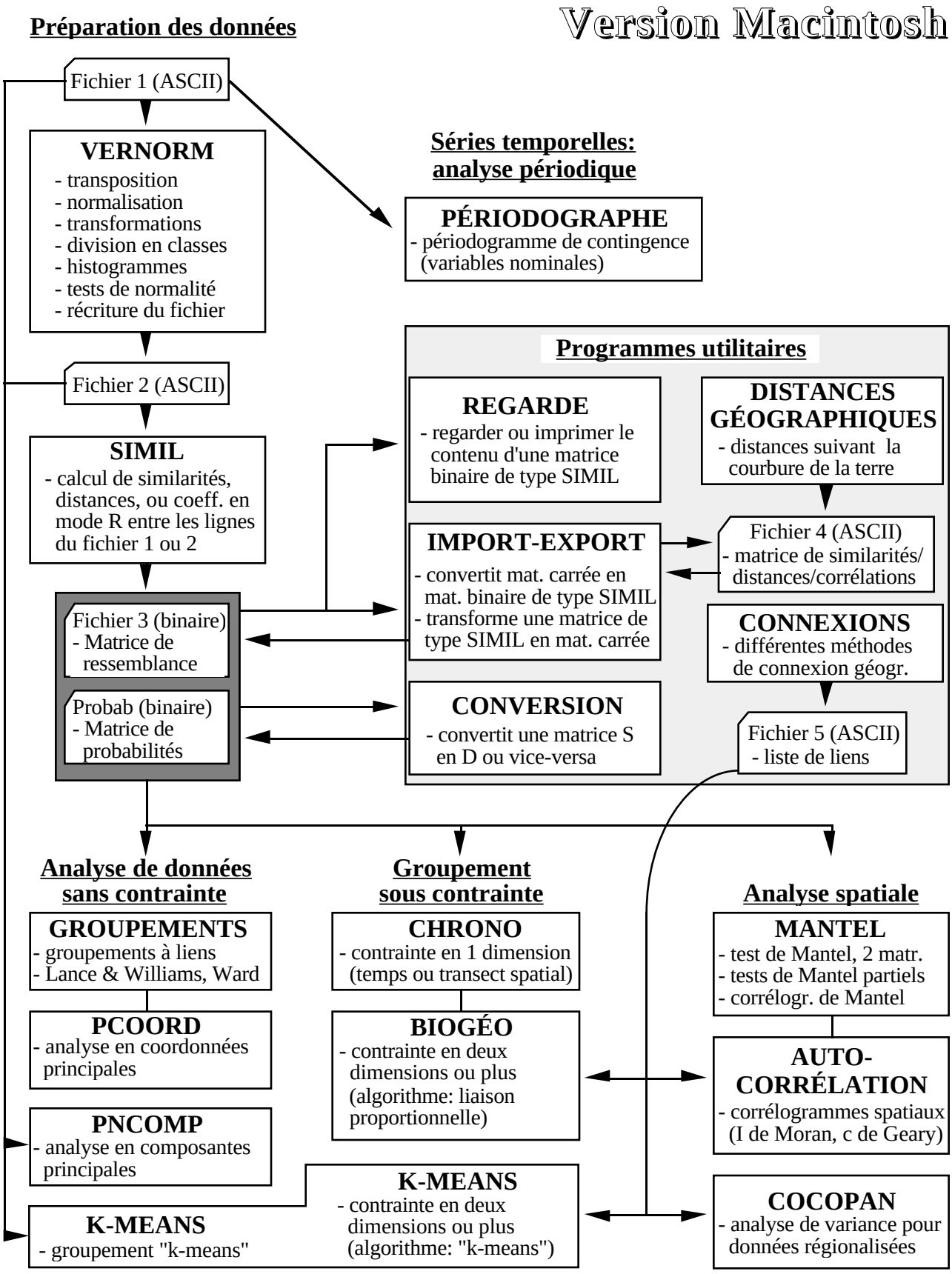
Ces programmes vous sont fournis sans aucune garantie implicite ou explicite de bon fonctionnement. Il s'agit de programmes mis au point dans le cadre de recherches universitaires. Cependant, si vous éprouviez des problèmes avec l'un ou l'autre des programmes de ce progiciel, nous serons heureux de tenter de vous dépanner et, du même coup, de régler ce problème pour l'ensemble des usagers de "R". Pour cela, il importe de nous fournir un maximum d'informations, et en particulier:

- La version du programme que vous utilisez (voir la ligne “Version” dans la fenêtre “Info”); date du programme (également dans la fenêtre “Info”) ou date à laquelle vous l'avez reçu.
- Les fichier(s) d'entrée; dans bien des cas, les problèmes qui nous sont soumis concernent simplement des erreurs de structure ou de contenu de ces fichiers. Sur Macintosh, les fichiers binaires de type SIMIL peuvent être compactés (par BINHEX ou STUFFIT) puis transmis par courrier électronique. Sur les grands ordinateurs IBM, les fichiers binaires de type SIMIL peuvent être transmis directement par courrier électronique.
- Les fichier(s) de sortie, incluant les messages que peuvent contenir ces fichiers.
- Tout autre message reçu à l'écran.

Veuillez soumettre ces informations à Alain Vaudor par courrier électronique, à l'adresse en couverture, ou à défaut, par courrier régulier (papier ou disquette).

Si vous installez ces programmes sur des machines différentes de celles sur lesquels ils ont été testés, il vous sera nécessaire de vérifier en détail le bon fonctionnement des programmes ainsi que la justesse des résultats. Il existe des différences de dialecte entre compilateurs PASCAL; de plus, les différences de longueur des mots-machine, ainsi que dans les valeurs minimum et maximum que peuvent prendre les nombres réels sur différentes machines, sont des sources potentielles de problèmes.

Version Macintosh



DESCRIPTION DES PROGRAMMES

*AUTOCORRÉLATION*Macintosh *ou AUTOCOR*CMS/VMS

Que fait AUTOCOR ?

Le programme AUTOCOR analyse l'autocorrélation spatiale d'une variable selon différents schémas de connexions et de distances entre les points. Cette méthode est strictement univariable; voir le programme MANTEL pour l'équivalent multivariable. L'autocorrélation est mesurée par les indices *I* de Moran et *c* de Geary, s'il s'agit de données quantitatives. Si les données sont ordinales ou nominales, les S.N.D. (*standard normal deviates*) sont calculés pour chaque classe de distance. Chaque valeur est accompagnée de la probabilité que celle-ci ne soit pas significativement différente de zéro (test unilatéral). L'interprétation des corrélogrammes est discutée par Legendre & Fortin (1989).

En version CMS ou VMS, ce même programme peut être employé pour produire une liste de paires d'objets (points) voisins dans une grille régulière (selon différentes stratégie de connexion), une triangulation de Delaunay ou un graphe de Gabriel. Ce fichier LIENS pourra servir par la suite de contrainte aux groupements réalisés par les programmes BIOGEO et KMEANS, ou en conjonction avec tout autre programme exigeant une liste de paires d'objets voisins, tel COCOPAN. Dans la version Macintosh, la fonction de fabrication du fichier de LIENS a été séparée et se trouve dans le programme CONNEXIONS. Enfin, ce programme peut aussi produire un fichier contenant une matrice triangulaire supérieure de classes de distance entre les objets. Ce fichier, appelé CLASSEF par défaut, est requis par le programme MANTEL pour calculer un corrélogramme multidimensionnel.

Fichiers d'entrée et de sortie

Les questions posées par le programme à propos des fichiers d'entrée et de sortie sont nombreuses et reflètent la multiplicité des options offertes. Lisez-les attentivement avant d'y répondre. Le programme requiert qu'on lui fournisse des informations quant (a) à la valeur que prend la variable en chaque point et (b) à la position relative des points. Il existe **cinq types de fichiers d'entrée** pour les versions VMS et CMS. Pour la version Macintosh, le fichier de données de type (2) n'est pas permis, car la fonction de fabrication des schémas de connexion et l'écriture du fichier de liens ont été transférées au nouveau programme CONNEXIONS.

(1) Liste des valeurs (Z)

Ce fichier d'entrée ne contient que les valeurs de la variable (appelée ici Z); il s'agit de nombre réels, ou encore d'entiers POSITIFS dans le cas d'une variable nominale. Dans ce fichier, on peut écrire les valeurs l'une à la suite de l'autre, séparées par un ou plusieurs espaces, suivant l'ordre des points, mais sans noms d'objets ou autres indicateurs; le programme assumera que le premier objet de la liste porte le numéro 1. La liste s'écrit de gauche à droite, en lignes successives, comme on lit une page de texte. Si on le désire, on peut n'inscrire qu'une seule valeur par ligne. Ce fichier de valeurs est le seul type qu'admet la version Macintosh; sa longueur est limitée à 16000 observations. Dans les versions CMS et VMS, on n'emploie ce fichier que dans le cas d'une grille régulière de points. Le schéma de connexions sera alors choisi par analogie avec le jeu d'échecs (voir Legendre & Legendre, 1984a, Tome 2, p. 257-259): mouvement de la tour (liens horizontaux et verticaux), du fou (diagonaux) ou de la reine (combinaison du fou et de la tour).

(2) Liste des coordonnées (X, Y) et des valeurs (Z)

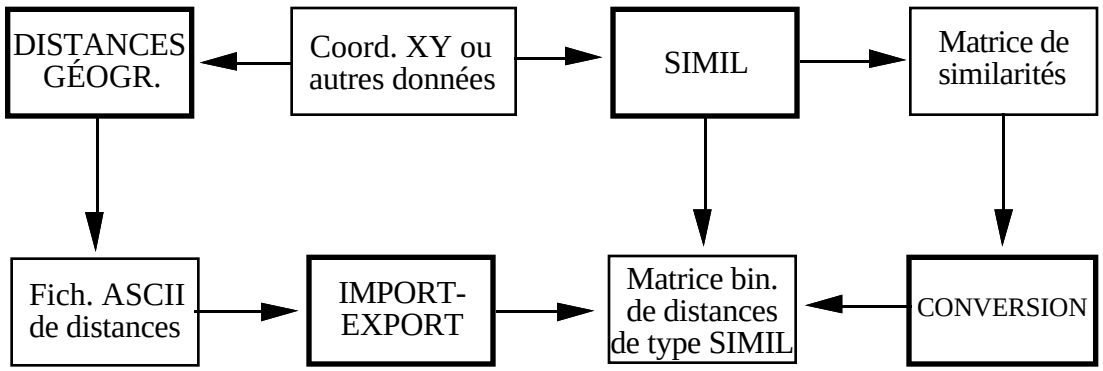
En versions CMS et VMS, lorsque les points ne forment pas une grille régulière, les coordonnées des points sont fournies dans le même fichier que les valeurs de la variable. Chaque

ligne de ce fichier doit donc contenir trois informations, comme suit:

Coordonnée en X Coordonnée en Y Valeur de la variable

Les coordonnées sont écrites sous la forme de nombres entiers ou réels (avec décimales) mais pas sous la forme de degrés-minutes-secondes. Elles sont lues en format libre; il n'est donc pas nécessaire de les disposer dans des colonnes précises. Comme pour les autres programmes CMS et VMS de ce progiciel, il faut écrire par exemple “0.376” et non pas “.376”. Voir p. 6 pour la version Macintosh.

(3) Matrice de distances



Les positions relatives des objets peuvent être décrites sous la forme d'une matrice de distances binaires calculée par SIMIL, ou transférée d’un autre programme par IMPORT (versions CMS et VMS) ou par IMPORT-EXPORT (version Macintosh). Le programme assumera qu'il s'agit d'une matrice de distances et **non** d'une matrice de similarités. Une matrice de similarités peut être aisément convertie en une matrice de distances à l'aide du programme utilitaire CONVERT. Le fichier à partir duquel SIMIL calcule la distance euclidienne (D01) doit doit contenir deux informations seulement:

Coordonnée en X Coordonnée en Y

Les coordonnées sont fournies sous la forme de nombres entiers ou de réels (*i.e.*, avec décimales) et non sous la forme de degrés-minutes-secondes. Ces données sont lues en format libre. L'avantage d'un tel fichier est que l'utilisateur peut choisir de calculer une distance autre que la distance géographique (euclidienne) entre objets. Le programme DIST peut aussi être mis à contribution pour calculer des distances suivant la courbure de la terre; ces distances se présentent sous la forme d'une matrice carrée (fichier ASCII) qu'il est aisé de convertir en format SIMIL à l'aide de l'utilitaire IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh).

(4) Matrice de classes de distances entre objets

Ce fichier en caractères lisibles (fichier ASCII et non binaire) peut représenter toute la matrice des distances déjà divisées en classes, ou encore la partie triangulaire supérieure seulement de cette matrice de distances, auquel cas elle se présente de la même façon que le fichier CLASSEF (type 8) décrit ci-dessous. Les classes de distance sont numérotées par les entiers successifs, débutant par 1. Cette matrice est de format $n \times n$ où n est le nombre de stations. Ce type de fichier permet à l'utilisateur qui le désire de soumettre une matrice qui n'est pas symétrique, c'est-à-dire une matrice dans laquelle la distance de *a* à *b* n'est pas nécessairement égale à la distance de *b* à *a*.

(5) Liste des liens entre les objets

Ce fichier en caractères lisibles (fichier ASCII et non binaire), fournit au programme une liste de

liens entre paires de points-objets. Chaque lien est représenté par une paire de numéros d'objets, écrits en format libre et séparés par au moins un espace. Ce fichier, qui peut être produit par ce même programme (version CMS/VMS) ou par le programme CONNEXIONS (version Macintosh), peut avoir par exemple l'apparence suivante (grille de 3 lignes et 4 colonnes, mouvement de la tour), où chaque paire de numéros représente un lien entre deux objets:

```
 1  2     2  3     3  4     5  6     6  7     7  8     9 10    10 11
11 12     5  1     6  2     7  3     8  4     9  5    10  6    11  7
12  8
```

Trois fichiers de sortie peuvent être créés par ce programme:

(6) Fichier des résultats contenant les statistiques du corrélogramme

Par défaut, ce fichier est appelé "SORTIE CORR A" dans la version CMS/VMS. Ce fichier ASCII diffère dans sa présentation selon que l'analyse porte sur des données quantitatives ou qualitatives (nominales). Un exemple pour chacun est donné plus bas. Lors de l'analyse de données quantitatives, les indices I de Moran et le c de Geary sont calculés pour chaque classe de distance d (Legendre & Legendre, 1984a, Tome 2, p. 258).

$$I(d) = [n \sum\sum w_{ij}(y_i - y\text{-}moy)(y_j - y\text{-}moy)]/[W \sum(y_i - y\text{-}moy)^2] \qquad \text{pour } i \neq j$$

$$c(d) = [(n\text{-}1) \sum\sum w_{ij}(y_i - y_j)^2]/[2W \sum(y_i - y\text{-}moy)^2] \qquad \text{pour } i \neq j$$

Les valeurs de la variable sont les y; y-moy désigne la moyenne de ces valeurs. Les w_{ij} prennent la valeur 1 quand la paire (i,j) appartient à la classe de distance d (celle pour laquelle on est en train de calculer la valeur du coefficient) et zéro dans les autres cas. W est la somme des valeurs w_{ij}, donc le nombre de paires dans toute la matrice carrée des distances entre points dont on a tenu compte pour calculer la valeur du coefficient pour la classe de distance sous considération. Le coefficient de Moran varie généralement de -1 à 1, mais il peut dans certains cas excéder -1 ou +1; les valeurs positives du I de Moran correspondent à de l'autocorrélation positive. Le coefficient de Gary varie de 0 à une valeur positive indéterminée qui n'excède que rarement 3 dans la plupart des cas réels; les valeurs de c inférieures à 1 correspondent à de l'autocorrélation positive.

Ces statistiques sont calculées pour chaque classe de distance disponible; chaque valeur est accompagnée de la probabilité que celle-ci ne soit pas significativement différente de zéro (test unilatéral). Les formules de calcul de l'erreur type de ces statistiques se trouvent dans Cliff & Ord (1981), Sokal & Oden (1978) et Legendre & Legendre (1984a). Les hypothèses sont les suivantes:

H_0: il n'y a pas d'autocorrélation spatiale. Les valeurs de la variable sont spatialement indépendantes les unes des autres. Chaque valeur du coefficient I est égale à $E(I) = -(n - 1)^{-1} \approx 0$, où $E(I)$ est l'espérance de I alors que n est le nombre de points d'observation; chaque valeur du coefficient c est égale à $E(c) = 1$.

H_1: il y a de l'autocorrélation significative. Les valeurs de la variable sont spatialement dépendantes les unes des autres. La valeur du coefficient I diffère significativement de $E(I) = -(n - 1)^{-1} \approx 0$; la valeur du coefficient c diffère significativement de $E(c) = 1$.

Tel que recommandé par Oden (1984), on pourra employer la correction de Bonferroni pour vérifier si le corrélogramme contient des valeurs significatives. Cette correction consiste à employer un niveau de signification *alpha*' = *alpha*/(le nombre de tests réalisés simultanément); par exemple, un corrélogramme de 5 classes de distance sera globalement significatif au niveau de 5% seulement s'il contient des valeurs significatives au niveau individuel de *alpha*' = 0.05/5 = 0.01.

Voici un exemple de fichier de sortie obtenu pour des données quantitatives, version Macintosh du programme; la sortie de la version CMS/VMS est virtuellement identique. Le corrélogramme correspondant est publié à la figure 3 de Legendre & Troussellier (1988).

```
P R O G R A M M E   A u t o C o r r é l a t i o n

Version Macintosh 1.0
Auteur: A. Vaudor


Matrice de distances:
    FICHIER D'ENTREE: XY, Thau
    TITRE: Distances géographiques, Thau (63 stations)
    DATE: 10/8/88
    FONCTION: D01
    Nombre d'objets : 63
    Nombre de descripteurs :  2

Classes équidistantes

   Classe  Limite sup.   Fréq.                  [données pour histogramme
        1   1.00518        97                  de fréquences des classes,
        2   2.01036       162                 dans la matrice triangulaire
        3   3.01553       250                              des distances]
      etc.           etc.
       17  17.08802         4

Fichier de données: CHLAtr
Nombre d'objets :  63

Option du mouvement:   Matrice SIMIL
Notes: Les probabilités sont plus significatives près de zéro.
       Les probabilités sont données à plus ou moins   0.00100

H0:             I = 0    I = 0            C = 1    C = 1
H1:             I > 0    I < 0            C < 1    C > 1
Dist.,I(Moran), p(H0),   p(H0), C(Geary),  p(H0),   p(H0),Paires
  1   0.4646    0.000            0.3355    0.000            194
  2   0.3833    0.000            0.4151    0.000            324
  3   0.3284    0.000            0.5352    0.000            500
  4   0.3382    0.000            0.5280    0.000            450
  5   0.2251    0.000            0.6708    0.000            484
  6   0.0773    0.101            0.8055    0.018            336
  7  -0.1109             0.121   1.0151              0.373   280
  8  -0.1992             0.011   1.1111              0.085   288
  9  -0.3517             0.000   1.3626              0.000   274
 10  -0.5869             0.000   1.7343              0.000   222
 11  -0.6228             0.000   1.8906              0.000   154
 12  -0.8550             0.000   2.2102              0.000   138
 13  -0.7459             0.000   2.4051              0.000   120
 14  -0.8355             0.000   2.5375              0.000    68
 15  -0.6122             0.001   2.4070              0.000    48
 16  -0.6631             0.023   2.4416              0.003    18
 17  -1.4980             0.001   3.3191              0.002     8

Total                                                      3906
```

En colonne 2 se trouve la valeur du *I* de Moran, et en colonne 5 la valeur du *c* de Geary, pour les différentes classes de distance (colonne 1). Les probabilités des tests unilatéraux pour le *I* de Moran sont présentées en colonnes 3 et 4; elles sont séparées en deux colonnes, selon que la valeur du coefficient est positive ou négative, de façon à en faciliter la lecture. Il en va de même pour les probabilités associées aux valeurs du *c* de Geary. Les hypothèses (H_0, H_1) sont spécifiées en haut de

ces colonnes. Par ailleurs, le nombre de paires de points correspondant à chaque classe de distance (cardinalité) forme la colonne de droite. Chaque nombre est le double de la valeur donnée dans l'histogramme de fréquence; c'est la valeur que l'on obtiendrait si on travaillait dans une matrice carrée, diagonale principale exclue, et non dans une matrice triangulaire de distances.

Dans la version Macintosh, le programme trace les corrélogrammes à l'écran et permet de les imprimer ou de les préserver dans des fichiers de type PICT. Un corrélogramme est un graphique dans lequel on porte les valeurs du coefficient d'autocorrélation spatiale (en ordonnée) en fonction des classes de distance (abscisse) (voir par exemple la figure 11.22 de Legendre et Legendre, 1984a). Voir aussi Legendre & Fortin (1989) pour l'interprétation des corrélogrammes spatiaux.

Pour les données nominales (qualitatives), ou encore pour les données ordinales traitées comme si elles étaient nominales, le programme calcule, pour chaque distance, les écarts normaux (S.N.D.: *standard normal deviates*) ainsi que les probabilités associées, pour chaque classe de distance et chaque paire d'états de la variable. La théorie relative à ces calculs est présentée par Sokal & Oden (1978), par Cliff & Ord (1981) ainsi que par Upton & Fingleton (1985). Voici un exemple de fichier de sortie obtenu pour des données nominales à 4 classes, obtenu à l'aide de la version CMS du programme. Peu de comparaisons sont significatives dans cet exemple.

```
A U T O C O R R E L A T I O N   S P A T I A L E

pour données quantitatives ou qualitatives.
Version IBM 2.0B
Auteur: Alain VAUDOR

Option du mouvement:   13

NOTE: Les probabilités les plus significatives sont près de zéro
Les probabilités sont imprimées à la précision de   0.00100

H0:                                S.N.D.=0,    S.N.D.=0
H1:                                S.N.D.>0,    S.N.D.<0
          CLASSES     S.N.D.         P(H0) ,       P(H0), PAIRES
DISTANCE  1                                                 312
           [1][1]     -0.272                        0.434
           [1][2]     -0.522                        0.301
           [1][3]     -1.068                        0.143
           [1][4]     -0.408                        0.342
           [2][2]      1.721         0.052
           [2][3]      0.889         0.187
           [2][4]     -1.687                        0.046
           [3][3] CARD. CLASSE [3]/NOBJ < 0.2 ou > 0.8
           [3][4]     -1.523                        0.064
           [4][4]      3.047         0.004
    [Total diff.]     -2.821                        0.002

DISTANCE  2                                                 586
           [1][1]     -2.204                        0.007
           [1][2]     -1.822                        0.034
           [1][3]     -0.246                        0.403
           [1][4]      2.001         0.023
           [2][2]      1.510         0.069
           [2][3]      1.485         0.069
           [2][4]      0.348         0.364
           [3][3] CARD. CLASSE [3]/NOBJ < 0.2 ou > 0.8
```

```
           [3][4]     -2.406                    0.008
           [4][4]     -0.082                    0.495
    [Total diff.]     -0.056                    0.478

DISTANCE  3                                          732
           etc.

DISTANCE  4                                          716
           etc.

DISTANCE  5                                          544
           etc.

DISTANCE  6                                          254
           etc.

DISTANCE  7                                           48
           etc.

TOTAL                                               3192
```

(7) Fichier des liens

Seules les versions CMS et VMS du programme peuvent produire ce fichier, qui est appelé "LIENS DATA A" par défaut. Ce fichier ASCII contient une liste de paires d'objets reconnus comme voisins par le schéma de connexion (options 1 à 13) utilisé lors de l'exécution du programme. Ce fichier LIENS pourra alors servir de contrainte aux groupements réalisés par les programmes BIOGEO et KMEANS, ou en conjonction avec tout autre programme exigeant une liste de paires d'objets voisins, tel COCOPAN. Dans la version Macintosh, ce fichier est produit par le programme CONNEXIONS. Un exemple de ce fichier est illustré en (5) ci-dessus. Il est à noter que l'utilisateur peut éditer ce fichier ASCII; il peut enlever des liens ou en ajouter, selon les besoins de son étude.

(8) Fichier contenant la matrice CLASSEF

Ce fichier ASCII est appelé "CLASSEF DATA A" par défaut dans la version CMS/VMS. Il contient la matrice triangulaire supérieure de classes de distance, permettant de calculer par la suite un corrélogramme de Mantel (voir la description du programme MANTEL).

Les options du programme

Les versions VMS et CMS offrent 16 options de calcul, numérotées de 0 à 15 (voir l'exemple, plus bas). On ne trouve que les options 0, 14 et 15 dans la version Macintosh. Ces options peuvent être regroupées dans les cinq catégories suivantes, en fonction des fichiers d'entrée dont on dispose.

(1) Option 0 — Matrice de distances de SIMIL

Pour cette option, deux fichiers d'entrée sont nécessaires: la Liste des valeurs (fichier de type 1) et la Matrice des distances (fichier de type 3) calculée à partir du coefficient de distance qui a été choisi par l'utilisateur (voir le tableau 4 pour la liste des coefficients du programme SIMIL). Pour cette option, il n'est pas nécessaire que les points soient disposés sur un grille régulière. Le programme pose les questions suivantes à l'utilisateur:

- “Classes équidistantes (0) ou équifréquentes (1) ?” — Les classes équidistantes sont de même

largeur d'intervalle de distances; les classes équifréquentes contiennent toutes le même nombre de paires, sauf dans les cas de données liées (distances égales) qui peuvent forcer certaines classes à contenir davantage de paires. On ne peut avoir l'un ET l'autre.

- "Nombre de classes ?" — L'usager doit déterminer combien de classes il désire obtenir.

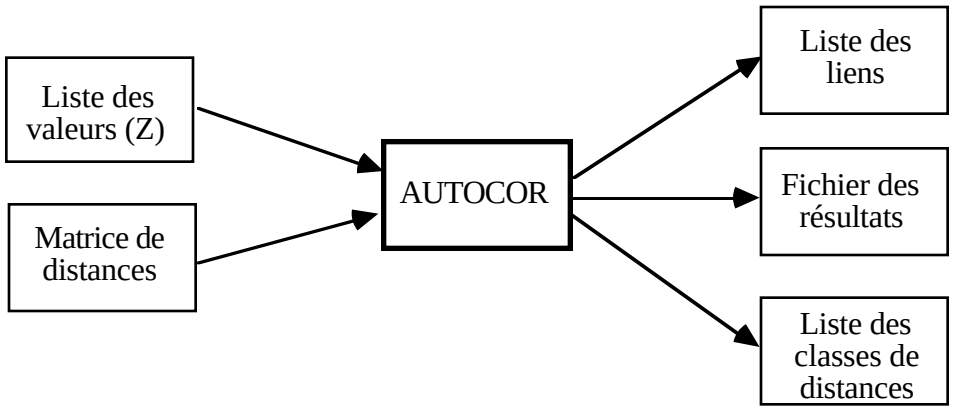
- "Désirez-vous voir l'histogramme ?" — Un histogramme permet d'apprécier la forme de la distribution des distances.

- "Préféreriez-vous un nombre/type différent de classes ?" — On a ici la possibilité de changer la division en classes en retournant aux deux premières questions.

- "Désirez-vous faire écrire la matrice CLASSEF des classes de distance, pour le corrélogramme de Mantel ?" — Voir la description de cette matrice au point (8) ci-dessus.

- "Désirez-vous écrire la liste des premiers liens sur le fichier "LIENS" ?" — Voir la description de ce fichier au point (7) ci-dessus.

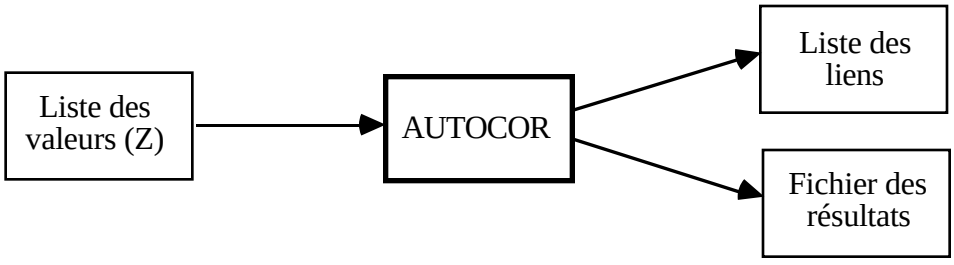
L'usager obtiendra les trois fichiers de sorties décrits aux points (6), (7) et (8) ci-dessus. Le fichier de sortie CLASSEF n'est disponible qu'à partir de cette option 0.



(2) Options 1 à 11 — Grille régulière

Ces options ne peuvent être utilisées que pour des points disposés selon une grille régulière; un seul fichier d'entrée est nécessaire: la Liste des valeurs (fichier de type 1). Ces options font référence pour la plupart à des types de connexion qui décrivent les mouvements du jeu d'échecs (réf : Legendre & Legendre, 1984, Tome 2, pp. 257-259), sauf pour ce qui est du calcul de la distance euclidienne entre les points de la grille. Le programme demandera quelle est la largeur et la hauteur de la grille qu'il devra confectionner. La distance entre deux points est le nombre minimum de liens qui les séparent.

L'usager peut obtenir le fichier des résultats (fichier de type 6) et le fichier des liens (type 7).



(3) Options 12 et 13 — Points disposés de façon irrégulière

Pour ces options, un seul fichier d'entrée est nécessaire, soit la Liste des coordonnées et des valeurs (fichier de type 2). Les connexions entre points sont alors calculées selon le graphique de Gabriel avec l'option 12 (Gabriel & Sokal, 1969) ou le système de triangulation de Delaunay avec l'option 13 (Dirichlet, 1850; Miles, 1970; Ripley, 1981; Watson, 1981; Upton & Fingleton, 1985; Isaaks & Srivastava, 1989). Voir le programme CONNEXIONS pour une description détaillée de ces méthodes. La distance entre deux points est calculée par le nombre minimum de liens qui les séparent.

Dans la triangulation de Delaunay (option 13), il y a deux façons d'imposer des "contraintes" à la formation de la triangulation plane. Voir la section portant sur la triangulation de Delaunay dans la description du programme CONNEXIONS. Rappelons qu'une "contrainte" est un ensemble de points supplémentaires, disposés à la périphérie des points-objets réels de l'étude. Dans la solution finale, tous les liens qui impliquent ces points supplémentaires sont éliminés; les points supplémentaires ont cependant, entre-temps, empêché la formation de longs liens entre les points périphériques du nuage de points, liens qui ne représentent pas des affinités réelles dans le cas des points périphériques distants mais sont simplement un effet de bordure de l'échantillonnage réalisé.

Deux méthodes sont disponibles dans le programme AUTOCOR pour imposer de telles "contraintes" à la formation de la triangulation. La question posée par le programme est la suivante:

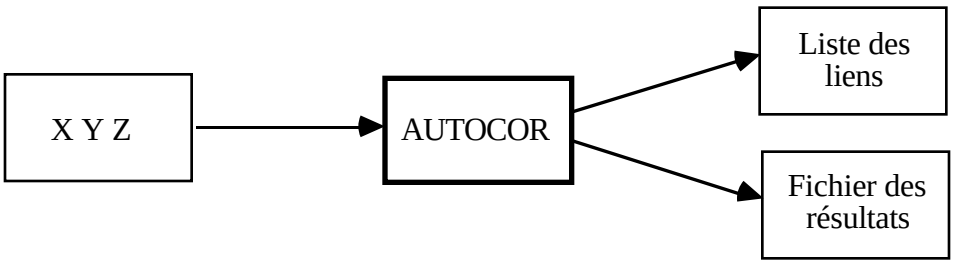
```
Nombre de points de contrainte? (-1 = contrainte rectangulaire)
```

1) Si on ne désire pas imposer de contrainte, on répond "0".

2) Si on désire imposer des contraintes rectangulaires, il n'est pas nécessaire de décrire ce cadre explicitement; il suffit de répondre "-1". Quatre points supplémentaires sont alors générés par le programme. Voir la description à la section portant sur la triangulation de Delaunay dans la description du programme CONNEXIONS.

3) Si l'usager désire imposer des "contraintes" en des endroits qu'il a lui-même judicieusement choisis, celles-ci doivent être décrites à la fin du fichier contenant la Liste des coordonnées et des valeurs. Chaque "contrainte" se présente sous la forme des coordonnées en X et en Y des deux points extrêmes du segment de droite formant la "contrainte"; donc, chaque "contrainte" est représentée par quatre chiffres: X_1 Y_1 X_2 Y_2. Le programme calcule alors les coordonnées du point milieu de ce segment et l'utilise comme "contrainte" dans les calculs subséquents. En réponse à la question, on indique le nombre de telles "contraintes" fournies au programme dans le fichier. **Cette façon de faire diffère de la procédure décrite dans le programme CONNEXIONS.**

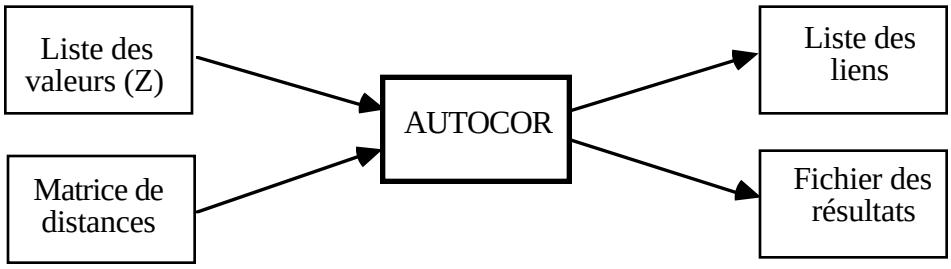
L'usager peut obtenir le fichier des résultats (fichier de type 6) et le fichier des liens (type 7).



(4) Option 14 — Votre propre matrice de classes de distance

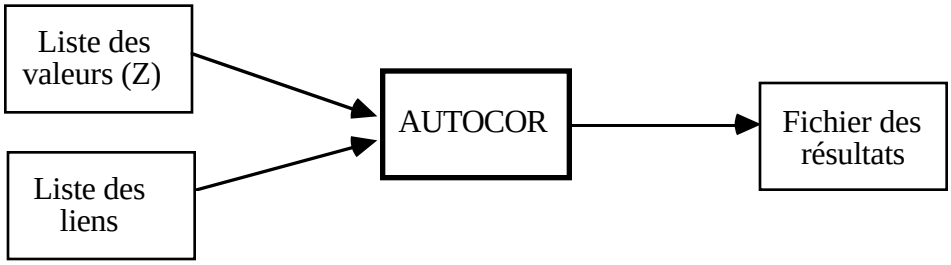
Deux fichiers d'entrée sont nécessaires: la Liste des valeurs (fichier de type 1) et la Matrice des classes de distance (fichier de type 4). La connexion entre les objets sera telle que décrite dans la matrice des classes de distance. Il n'est pas nécessaire que les points forment une grille régulière. En

sortie, l'usager peut obtenir le fichier des résultats (fichier de type 6) et le fichier des liens (type 7).



(5) Option 15 — Votre propre liste de liens

Pour cette option, deux fichiers d'entrée sont nécessaires: la Liste des valeurs (fichier de type 1) et la Liste des liens (fichier de type 5). La connexion entre les objets sera telle que décrite dans la liste des liens. Il n'est pas nécessaire que les points forment une grille régulière. La distance entre deux points est mesurée par le nombre minimum de liens qui les séparent. L'usager ne pourra obtenir en sortie que le fichier des résultats (de type 6).



Les questions du programme

L'exemple ci-dessous montre le dialogue que propose le programme en version CMS/VMS; les réponses données par l'usager sont soulignées et en caractère gras. Les questions posées par la version Macintosh sont essentiellement les mêmes, quoique leur formulation pourra parfois différer légèrement. Les explications qui suivent correspondent aux numéros en marge gauche de l'exemple.

(1) L'usager déclare d'abord que ses données ne sont pas nominales.

(2) Entre les points d'observation, une triangulation de Delaunay sera calculée (option 13); la distance entre les points est le *nombre de liens* formant le plus court chemin entre deux points, en suivant les liens de la triangulation.

(3) Il y a 57 points dans le fichier à l'étude. Si on avait décrit une "contrainte" par une liste de points supplémentaires, ces points ne seraient pas comptés en réponse à cette question.

(4) On opte pour une "contrainte" rectangulaire (voir ci-dessus).

(5) Dans le cas d'une grille régulière ou partiellement régulière, il pourra arriver que deux solutions soient totalement équivalentes et que deux traits se croisent. L'usager pourra décider soit de garder ces deux traits équivalents, soit d'éliminer l'un des deux. Une telle situation ne peut se produire avec l'algorithme mis en oeuvre dans la version Macintosh (programme CONNEXIONS).

(6) L'utilisateur demande que la liste des liens soit inscrite dans le fichier LIENS, pour usage ultérieur.

Exemple

```
Analyse de l'AUTOCORRELATION SPATIALE.

Pour toutes les options sauf 12 et 13, vous aurez besoin d'un
fichier de VALEURS. Pour les options 12 et 13, vous aurez
besoin d'un fichier de COORDONNEES contenant aussi, en troisieme
position, les VALEURS de la variable.

Pour l'option 13 (Delaunay), si vous desirez imposer des segments
de contrainte, ceux-ci doivent apparaitre dans ce meme fichier,
a la fin de la liste des points-objets, sous la forme de 2 points
(4 coordonnees) decrivant chaque segment.

Quel est le nom de ce fichier? (Par defaut: "... data a")
*** Vous DEVEZ fournir un fichier de donnees, meme si vous
*** ne desirez que la liste des liens et n'etes pas interesse
*** au correlogramme.
fichier  data  a

Pour l'option 0, vous aurez besoin d'une matrice binaire
de DISTANCES, produite par SIMIL ou IMPORT.
Assurez-vous qu'il ne s'agit PAS d'une matrice de similarites.

Quel est le nom du fichier contenant cette matrice s'il y a lieu?
(Par defaut: "... data a")


Pour l'option 14, quel est le nom de la matrice de classes
de distance, s'il y a lieu (carree ou triangulaire superieure) ?
(Par defaut: "... data a")


Pour l'option 15, quel est le nom du fichier de liens que vous
avez prepare, s'il y a lieu? (Par defaut: "... data a")


Quel nom desirez-vous donner au fichier de sortie, contenant
le correlogramme? (Par defaut: "Sortie corr a")


Quel nom desirez-vous donner au fichier de LIENS produit par ce
programme, s'il y a lieu? (Par defaut: "Liens data a")


Quel nom desirez-vous donner au fichier contenant la matrice
CLASSEF (matrice triangulaire superieure de classes de
distance, permettant de calculer par la suite un correlogramme
de Mantel), s'il y a lieu ?  (Par defaut: "Classef data a")
```

```
     A U T O C O R R E L A T I O N   S P A T I A L E

     pour donnees quantitatives ou qualitatives.
     Version IBM 2.0B
     Auteur: Alain VAUDOR

     Votre fichier de donnees est-il deja en classes ?
     Autrement dit, desirez-vous analyser des DONNEES QUALITATIVES ?
(1)  n

     OPTIONS:
       0: Matrice de distances de SIMIL (Fichier "ENTREEB")

      MOVEMENTS DANS UNE SEULE DIRECTION:
       1: Mouvement horizontal (Lignes)
       2: Mouvement vertical (Colonnes)
       3: Mouvement diagonal (pente positive)
       4: Mouvement diagonal (pente negative)

      JEU D'ECHECS, MOUVEMENTS DIRECT SEULEMENT:
       5: Mouvement de la tour
       6: Mouvement du fou
       7: Mouvement de la reine

      JEU D'ECHECS, MOUVEMENTS DIRECTS ET INDIRECTS:
       8: Mouvement de la tour
       9: Mouvement du fou
      10: Mouvement de la reine

      11: Distance euclidienne, points en grille reguliere

      POINTS DISPOSES DE FACON IRREGULIERE:
      12: Graphique de Gabriel
      13: Triangulation de Delaunay
      14: Votre propre matrice de classes de distance
      15: Votre propre liste de liens (attacher fichier "LIENS")

(2)  13

     Nombre total de points ?
(3)  57

     Nombre de points de contrainte? (-1 = contrainte rectangulaire)
(4)  -1

     Elimination des traits qui se coupent?
(5)  o

     Desirez-vous ecrire la liste des premiers liens sur le fichier "LIENS" ?
(6)  o

     ***  312 liens ont ete ecrits sur le fichier de LIENS ***
     Fin du programme.
```

BIOGÉO

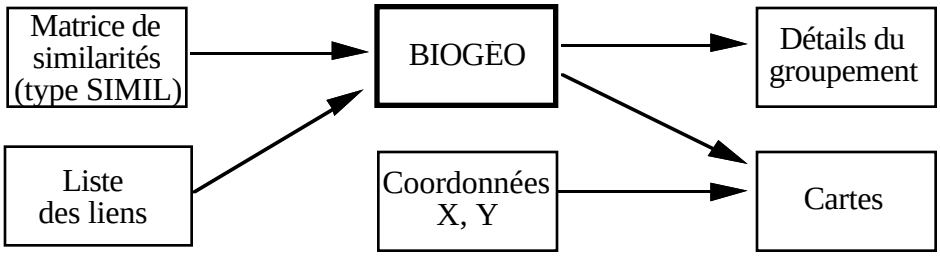
Que fait BIOGÉO ?

Ce programme calcule un groupement agglomératif avec contrainte de contiguïté spatiale, tel que proposé par Legendre & Legendre (1984b), et présente les résultats sous forme d'une série de cartes, une pour chaque niveau de groupement. Puisque le groupement est basé sur une matrice de similarités et que cette matrice est le plus souvent calculée à partir d'un grand nombre de descripteurs, cette méthode peut donc être considérée comme une méthode de cartographie multidimensionnelle.

Le groupement agglomératif procède selon un algorithme à liaison proportionnelle; un autre programme de ce progiciel, K-MEANS, permet de réaliser du groupement sous contrainte à l'aide d'un algorithme non-hiérarchique. La connexité est fixée par l'usager entre 0 (groupement à liens simples) et 1 (groupement à liens complets). Legendre (1987) a montré la stabilité des résultats du groupement avec contrainte à travers une large gamme de valeurs de connexité.

Si les dimensions actuelles du programme (version CMS ou VMS) sont insuffisantes, celles-ci peuvent aisément être modifiées en changeant la valeur des paramètres en début de programme et en le recompilant. Tel est également le cas de tous les autres programmes de ce progiciel. Dans la version Macintosh, une limite du programme impose de ne jamais avoir plus de 150 groupes simultanément. Des problèmes comportant plus de 1000 objets ont été traités par ce programme; il peut être nécessaire, dans de tels cas, de demander plus de mémoire que la quantité attribuée par défaut aux usagers.

Fichiers d'entrée et de sortie



(1) Fichier de similarités

Le fichier de similarités produit par les programmes SIMIL, IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh), qui décrit les relations de ressemblance entre points, est toujours nécessaire à ce programme. Une matrice de distances devra être convertie en matrice de similarités par le programme CONVERSION avant d'être utilisable par BIOGÉO.

(2) Fichier de liens

Les relations spatiales entre les points doivent être fournies au programme sous la forme d'une liste de liens (fichier LIENS, en ASCII et non en binaire). Chaque lien est représenté par une paire de numéros d'objets, écrits en format libre et séparés par au moins un espace. Ce fichier, qui peut être produit par les programmes AUTOCOR (version CMS/VMS) ou CONNEXIONS (version Macintosh), peut avoir par exemple l'apparence suivante (grille de 12 points disposés en 3 lignes et 4 colonnes, mouvement de la tour), où chaque paire de numéros représente un lien entre deux objets:

```
 1  2     2  3     3  4     5  6     6  7     7  8     9 10    10 11
11 12     5  1     6  2     7  3     8  4     9  5    10  6    11  7
12  8
```

Ce fichier peut être modifié à l'aide d'un éditeur ASCII si on désire ajouter ou retrancher des liens de la liste. Le fichier peut également être entièrement écrit à l'aide de l'éditeur; on peut ainsi, par exemple, fournir la liste des premiers ET des deuxièmes voisins de chaque point, ou toute autre combinaison jugée intéressante en fonction de la problématique de l'étude. Si la liste inclut toutes les paires possibles de points-objets, le groupement devient sans contrainte; cette option est disponible dans la version Macintosh. Il peut être intéressant d'utiliser BIOGEO de cette façon, puisqu'on peut ainsi obtenir une carte pour chaque étape du groupement.

Avant de faire démarrer le programme, assurez-vous que vous connaissez le nombre de liens (paires de points) qui doivent être lus par le programme dans le fichier de LIENS. Suggestion: intégrez ce nombre au nom du fichier.

(3) Fichier de coordonnées spatiales (X, Y)

Si on désire demander au programme de tracer les cartes correspondant à chaque niveau de groupement (option du programme), il faut lui fournir un fichier contenant les coordonnées des points à analyser. C'est à partir de ces coordonnées que la position des points sera établie sur ces cartes. Les coordonnées sont fournies en format lisible (non en binaire) sous la forme d'entiers ou de nombres réels en degrés décimaux. Les coordonnées **ne doivent pas** être en degrés-minutes-secondes. Le nombre de coordonnées dans ce fichier doit correspondre au nombre d'objets. Avec les versions CMS et VMS, n'oubliez pas de mettre un zéro avant le point décimal ("0.376" et non pas ".376").

Pour certaines représentations didactiques, on pourra fournir dans ce fichier des coordonnées qui ne correspondent pas exactement aux positions géographiques. Par exemple, pour analyser d'un seul bloc des échantillonnages répétés d'un même territoire au cours du temps, on pourra prévoir la position des objets de l'étude de façon à ce que chaque tranche de temps forme une partie séparée de l'image finale. Les coordonnées fournies dans ce fichier ne servent qu'à l'illustration; les relations spatiales ou spatio-temporelles qui sont tenues en compte lors du groupement sont uniquement celles que contient le fichier de liens.

(4) Fichier des similarités triées

Dans les versions CMS et VMS, il est possible de conserver le fichier des similarités triées pour un calcul subséquent. Cette option est particulièrement intéressante lorsqu'on désire étudier les résultats obtenus avec plusieurs valeurs différentes de connexité, alors que la matrice de similarité est grande et donc longue à trier.

(5) Fichier de résultats

En versions CMS et VMS, la seule sortie de BIOGÉO est le fichier de résultats contenant les détails du groupement et les cartes. Le nombre de cartes disponibles correspond au nombre d'étapes du groupement, soit n - 1. L'usager peut cependant préférer ne pas faire inscrire toutes les cartes dans le fichier, les premières cartes, qui correspondent à des niveaux élevés de similarité, étant souvent peu informatives; on peut donc indiquer combien des dernières cartes on désire obtenir dans le fichier de résultats. Voir la section "Contenu du fichier de résultats" pour plus de détails sur le fichier de sortie.

Dans la version Macintosh, la fonction cartographique est séparée du fichier détaillant les résultats du groupement. Le fichier contenant le détail des étapes de groupement est optionnel. Par ailleurs, si on désire obtenir les cartes, celles-ci sont présentées à l'écran une à une. L'usager peut choisir la carte désirée soit par son niveau de similarité, soit à l'aide d'un curseur qui indique le nombre de groupes obtenu à chaque niveau de similarité (le début du groupement, et donc les similarités élevées, sont au bas de l'écran); on fixe la position du curseur à un niveau de similarité donné, connaissant le nombre de groupes présents à ce niveau, et on clique la souris. Voir également les autres options dans le menu déroulant "Choix de cartes" du programme. Notez que plusieurs

étant illustré par une carte séparée. Sur la carte qui apparaît à l'écran, les membres d'un même groupe sont entourés d'un trait formant une enveloppe, si la situation le permet; les enveloppes peuvent être prolongées par des traits dendritiques au besoin. Si on n'est pas certain de la séparation des groupes dans une portion de l'image, on peut agrandir n'importe quelle partie de celle-ci en traçant un rectangle autour de ladite portion à l'aide de la souris. Une nouvelle section de la partie agrandie peut à son tour être agrandie; la commande "Terminer" du menu déroulant permet de revenir à la carte précédente. L'usager pourra demander d'imprimer les cartes de son choix ou de les conserver dans des fichiers de type PICT; les cartes sont identifiées par un titre et un niveau de similarité de groupement.

Les options du programme

Les options du programme sont les suivantes. Les numéros se réfèrent aux numéros en marge gauche de l'exemple ci-dessous.

- Le choix du niveau de connexité (*Co*) du groupement agglomératif à liaison proportionnelle (4).
- La possibilité d'obtenir les cartes, ou non (2 et 5).
- La possibilité, en version Macintosh, d'obtenir ou non le détail des groupes formés à chaque niveau.
- La possibilité, en versions VMS et CMS, de conserver le fichier des similarités triées (1 et 3).

Exemple

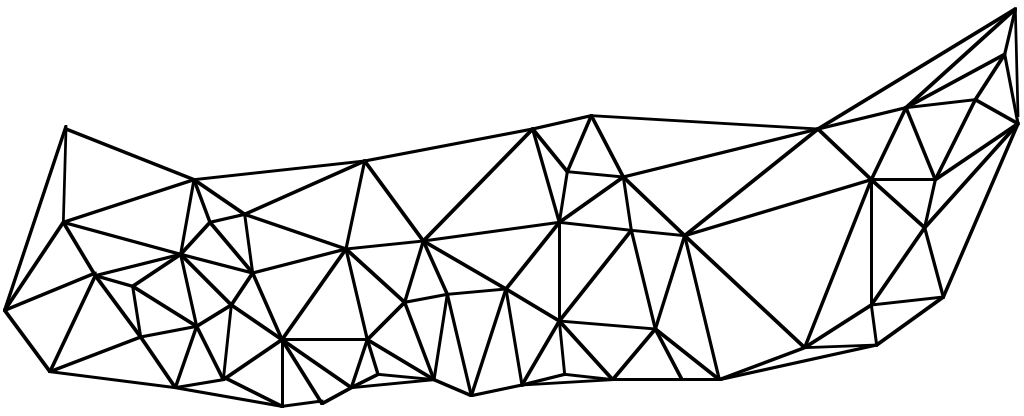
L'exemple ci-dessous illustre l'utilisation du programme pour calculer un groupement sous contrainte de contiguïté spatiale. Le fichier d'appel, dont le dialogue forme la première partie de l'exemple, demande le nom des divers fichiers. Cet exemple a été réalisé sous CMS. Les questions posées par la version Macintosh sont essentiellement les mêmes, quoique leur formulation peut parfois différer légèrement. Le premier point à signaler concerne le fichier des similarités triées (1): on donne un nom en réponse à cette question si on désire conserver le fichier trié, ou encore si, ayant conservé un tel fichier au cours d'une passe précédente, on désire maintenant l'utiliser (auquel cas il faut également répondre "oui" en (3) en réponse à la question du programme). Le second point est que le fichier des coordonnées est optionnel (2); il n'est requis que si l'on désire les cartes que ce programme peut produire (auquel cas il faut également répondre "oui" en (5) à la question du programme). Pour le traçage des cartes en versions CMS et VMS, la première colonne du fichier de coordonnées sera l'abscisse (valeurs croissantes de gauche à droite) et la seconde colonne l'ordonnée (valeurs croissantes du bas vers le haut); l'usager doit déterminer quelle largeur aura sa carte, en réponse à la question (6). En version Macintosh, la coordonnée présentant la plus grande plage de variation est toujours l'abscisse, avec rotation de l'image au besoin de façon à occuper l'écran au mieux.

```
     BIOGEO: Groupement sous contrainte de contiguite spatiale.

     Quel est le nom du fichier contenant la matrice de SIMILARITES
     de type SIMIL? (Par defaut: "... data a")
     (Il faut fournir ce fichier meme si vous fournissez un fichier
     de similarites triees, en reponse a la question suivante.)
     fichier  s16  a

(1)  Desirez-vous conserver le fichier de similarites triees
     pour utilisation future?
     Ou encore, possedez-vous deja ce fichier? Dans l'un ou
     l'autre cas, quel est son nom? (Par defaut: "FICHTRI data a")
     fichier  fichtri  a

     Quel est le nom du fichier contenant la liste des LIENS DE
```

```
     PROXIMITE a employer comme contrainte du groupement?
     (Par defaut: "... data a")
     fichier liens146 a

(2)  Quel est le nom du fichier des COORDONNEES des localites,
     s'il y a lieu? (Par defaut: "... data a")
     fichier coord a

     Quel nom doit recevoir le fichier de sortie de BIOGEO?
     (Par defaut: "CARTES BIOGEO a")
     fichier cartes a

      P r o g r a m m e  B I O G E O

      Auteur: A. Vaudor

(3)   Avez-vous fourni un fichier de similarites deja triees (FICHTRI) ?
      (O ou N)
     n
      Titre de ce travail
     Groupement sous contrainte spatiale
      Nombre de paires dans le fichier de liens?
     146
(4)   Connexite desiree (Max: quatre chiffres significatifs)
     1.0
      Il y a 56 etapes de groupement.
      Combien des dernieres etapes vous interessent?
     20
(5)   Desirez-vous les cartes? (O ou N)
     o
(6)   Largeur des cartes (en caracteres, sans compter le cadre):
     60
      Fin du programme.
```

Contenu du fichier de résultats

Le fichier présenté ci-dessous est une sortie du programme en version CMS. Pour chaque niveau de groupement, les cartes ont été demandées, en plus du détail du groupement. La connexité du groupement à liaison proportionnelle a été fixée à *Co* = 1.0. Les relations de voisinage spatial entre les points, décrites par le fichier des liens, sont les suivantes (image produite par CONNEXIONS):



```
B I O G E O : Groupement sous contrainte spatiale

Auteur: A. Vaudor
```

Dans la liste des 57 objets, ci-dessous, chaque objet est identifié par le numéro de son groupe. Les numéros de groupes ne sont pas nécessairement séquentiels. Les objets non encore groupés reçoivent un zéro.

```
Niveau:              1.00000
Connexité:           1.00000
Nombre de groupes:   9

  1   0   0   1   1   0   1   1   1   5   5   0   5   1   1  16  16  16
  1   0   1   1   1   1   1   0   2   2   2   2   2   2   2   6   2   2
  2   6   6   0   4   4   4   4   0   0   4  12  12   0  13  13  13  12
 15  15  13
Nombre de localites groupées:  47
------------------------------------------------------------
!                                                  1       !
!                                                          !
!                                                1         !
!   +                        2  2                      1   !
!          %       2             1           1  1          !
!   %        %                1   1              1         !
!         *        2  2              1                     !
!     * *     4      2  2  1                  5   5        !
!+      4      4    6             1       5                !
!  *        4      62  2   =  = =   1 1                    !
!         4      6       2                                 !
------------------------------------------------------------
```

Après le no 9, les symboles utilisés dans la carte n'ont plus de rapport avec le numéro du groupe.

```
Niveau:              0.12500
Connexité:           1.00000
Nombre de groupes:   4

  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
  1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   4   2   2
  2   4   4   4   4   4   4   4   3   3   4   2   2   2   4   4   4   2
  2   2   4
Nombre de localites groupées:  57
------------------------------------------------------------
!                                                  1       !
!                                                1         !
!                                              1   1       !
!   2                        2  2           1        1     !
!          2       2           1  1            1  1        !
!   2       22                 1   1              1        !
!         4        2  2              1                     !
!     4 4    34      2  2  1                  1   1        !
!2       4  3   4    4         1    1       1   1          !
!  4        4      42  2   1  1 1   1 1                    !
!         4    4 4       2                                 !
------------------------------------------------------------
```

Les cartes produites par la version Macintosh sont de meilleure qualité graphique (voir section K-MEANS). Les objets y sont représentés par leur numéro d'ordre dans le fichier d'entrée. Les groupes sont matérialisés par des enveloppes entourant les points membres d'un même groupe.

CHRONO

Que fait CHRONO ?

Ce programme calcule le groupement chronologique proposé par Legendre, Dallot & Legendre (1985). Cette méthode de groupement, d'abord décrite pour les séries temporelles de données multivariables, peut aussi être employée pour l'analyse des séries spatiales (Galzin & Legendre, 1987). Le groupement non-hiérarchique procède selon un algorithme agglomératif à liaison proportionnelle, dont le degré de connexité (*Co*) est fixé par l'usager en réponse à une question du programme; c'est le test de signification, décrit au paragraphe suivant, qui rend le résultat non-hiérarchique. La contrainte de contiguïté temporelle ou spatiale imposée au groupement signifie que seuls les objets ou les groupes d'objets adjacents le long de la série peuvent se grouper. Fait à noter, il est peu probable que de changer la connexité change de façon notoire les résultats du groupement, comme on peut le voir dans les exemples de la publication de Legendre, Dallot & Legendre (1985).

À chaque étape du groupement agglomératif, un test par permutation est réalisé pour décider si on doit, ou non, fusionner les deux groupes dont la fusion est proposée par l'algorithme agglomératif. L'hypothèse nulle de ce test est décrite explicitement dans la liste de sortie des versions CMS et VMS:

```
H  est  la probabilite  que  l'hypothese  principale  soit
vraie.  Selon celle-ci,  les deux groupes  soumis  au test
sont un artefact  et devraient  etre fusionnes  en un seul
groupe. La fusion est accomplie si H est plus eleve que le
seuil de probabilite ALPHA etabli plus haut par l'usager.
```

En réponse à une question du programme, l'usager doit fixer lui-même le niveau *alpha* de rejet de l'hypothèse nulle (souvent 0.01, 0.05 ou 0.10; il est cependant possible de tester à un niveau plus élevé pour identifier les singletons — voir ci-dessous, ainsi que l'exemple). Il faut réaliser qu'il ne s'agit pas d'un véritable test d'hypothèse statistique, les données servant au test étant les mêmes que celles qui ont servi à générer l'hypothèse de division en groupes. Des simulations, décrites dans la référence principale, ont cependant montré que pour des données aléatoires, la probabilité que ce test produise un résultat significatif est bien égal à *alpha*.

Le programme permet d'identifier les *singletons*, ou prélèvements aberrants se trouvant le long de la série. La présence d'un singleton peut empêcher la formation d'un groupe qui aurait inclus des objets situés de part et d'autre du prélèvement aberrant. Trois raisons au moins peuvent entraîner la formation de prélèvements aberrants: (1) des événements aléatoires, tels que des strates modifiées dans une carotte de sédiments, ou encore des mouvements de masses d'eau lors d'un échantillonnage répété au cours du temps à une station fixe en milieu aquatique; (2) des problèmes d'échantillonnage ou de préservation des échantillons; (3) des variations stochastiques extrêmes, qui font que l'hypothèse nulle sera rejetée alors qu'il n'y a pas eu brisure dans la succession (erreur de type II).

Si l'usager demande d'identifier les singletons, ceux-ci seront éliminés de la série et le groupement sera repris depuis le départ (voir l'exemple); font exception à cette règle les singletons situés en bout de série (début ou fin), puisque aucun groupe n'est bloqué par leur présence. Il est peu probable que l'on réussisse à identifier des singletons si le niveau *alpha* est faible (moins de 10 %), parce qu'il devient difficile, lors du test d'un seul objet contre p objets, d'obtenir une valeur inférieure à celles de la première colonne du tableau 1. Enfin, si un objet a une similarité de zéro avec tous ses voisins immédiats, le groupement agglomératif ne se rend pas jusqu'au niveau $S = 0$ pour tenter de l'inclure dans un groupe; un tel objet non groupé est représenté par un tiret (-) dans le groupement final, ou encore par un carré blanc dans le dessin de la version Macintosh. L'usager devra vérifier les données de tout objet ainsi identifié; il est recommandé de l'éliminer de l'analyse, s'il s'agit d'un objet aberrant ou exceptionnel ayant une similarité nulle avec ses voisins, au cas où sa présence dans la série ait interrompu la formation d'un groupe englobant des objets situés de part et d'autre.

Tableau 1 — Les plus faibles probabilités de fusion possibles pour deux groupes de taille p_1 et p_2 respectivement (excepté dans des cas d'égalité des valeurs de similarité). Tiré de Legendre *et al.* (1985), Tableau C1.

	p_1				
p_2	1	2	3	4	5
2	0.66667	0.33333			
3	0.25000	0.10000	0.10000		
4	0.20000	0.06667	0.02857	0.02857	
5	0.16667	0.04762	0.01786	0.00794	0.00794
6	0.14286	0.03571	0.01190	0.00476	0.00217
7	0.12500	0.02778	0.00833	0.00303	0.00126
8	0.11111	0.02222	0.00666	0.00202	0.00078
9	0.10000	0.01818	0.00455	0.00140	0.00050
10	0.09091	0.01515	0.00350	0.00100	0.00033
11	0.08333	0.01282	0.00275	0.00073	0.00023
12	0.07692	0.01099	0.00220	0.00055	0.00016
13	0.07143	0.00952	0.00179	0.00042	0.00012
14	0.06667	0.00833	0.00147	0.00033	0.00009
15	0.06250	0.00735	0.00123	0.00026	0.00006
16	0.05882	0.00654	0.00103	0.00021	0.00005
17	0.05556	0.00585	0.00088	0.00017	0.00004
18	0.05263	0.00526	0.00075	0.00014	0.00003
19	0.05000	0.00476	0.00065	0.00011	0.00002
20	0.04762	0.00433	0.00056	0.00009	0.00002

Fichiers d'entrée et de sortie

(1) Le fichier d'entrée

Le fichier d'entrée doit impérativement être un fichier de similarités, et NON PAS de distances, produit par le programme SIMIL, ou encore par IMPORT (en versions CMS et VMS) ou IMPORT-EXPORT (en version Macintosh). Une matrice de distances peut être aisément convertie en une matrice de similarités à l'aide de l'utilitaire CONVERSION (CONVERT en version VMS/CMS). Le programme assume que l'ordre chronologique ou temporel est le même que l'ordre des objets.

(2) Les résultats

Les résultats du calcul, qui sont présentés à l'écran (versions CMS et VMS) ou à la fois dans un fichier et à l'écran (version Macintosh), montrent d'abord le groupement. Quoique la méthode pour y arriver soit hiérarchique, le résultat final est non-hiérarchique. Ce résultat est illustré par le dessin à l'écran dans la version Macintosh. Il est également présenté à la *dernière ligne* de la liste illustrant les étapes du groupement (à l'écran pour les versions CMS et VMS; dans un fichier pour la version

Macintosh); les lignes qui précèdent, peu informatives, ne sont présentées que pour indiquer à l'usager que le programme est en train de travailler pour lui. Seule la dernière ligne de cette liste est donc à conserver et à reproduire dans les publications.

Des tests *a posteriori* peuvent être réalisés, qui permettent de procéder à l'expansion de chaque groupe à tour de rôle, en supposant que les autres groupes n'existent pas et que leurs objets sont encore des points-observations isolés; cette expansion des groupes permet de déterminer si les groupes formés lors du groupement sont séparés de façon brusque les uns des autres (succession par sauts), ou si au contraire la transition entre eux est douce (succession graduelle). D'autres tests *a posteriori* permettent de connaître les relations entre groupes distants et de déterminer si certains seraient semblables (on se référera à l'hypothèse nulle pour comprendre dans quel sens interpréter ces tests; voir aussi l'exemple ci-dessous). Le programme fait de même avec les singletons, tentant de déterminer s'ils ressemblent à l'un ou l'autre des groupes distants. Dans ces tests *a posteriori*, plusieurs fusions entre petits groupes seront réalisées simplement à cause du fait qu'il est impossible aux tests de prendre des valeurs de probabilités inférieures aux valeurs minimales décrites au tableau 1 — spécialement si le seuil *alpha* fixé en début de groupement est faible.

On notera que les tests *a posteriori*, et en particulier l'expansion des groupes, sont coûteux en temps de calcul. On ne les réalise habituellement pas au cours des analyses exploratoires d'un fichier de données; on attendra plutôt que la combinaison la plus informative des paramètres du programme (connexité et niveau *alpha*) ait été déterminée. Dans les versions pour grands ordinateurs, si on désire conserver ces résultats et les faire imprimer, il faut les faire inscrire dans un fichier de "trace de la console" (version CMS) tel qu'expliqué à la page 2 du présent document.

Les options du programme

Les options du programme sont les suivantes. Les numéros se réfèrent aux numéros en marge gauche de l'exemple ci-dessous.

- Le choix du niveau de connexité (*Co*) du groupement agglomératif à liaison proportionnelle (1).
- Le choix du niveau de signification *alpha* du test par permutation (2).
- La possibilité d'éliminer les singletons (3).
- Les tests *a posteriori*: expansion des groupes (4), tests entre les groupes distants (5).

Exemple

L'exemple ci-dessous illustre l'utilisation du programme pour calculer un groupement sous contrainte de contiguïté en une dimension (spatiale, dans ce cas). Le fichier de données représente un transect spatial (*i.e.*, une radiale) de 24 stations où 41 espèces ont été identifiées. Le coefficient de similarité de Steinhaus (S17) a été employé pour comparer les stations. Dans cet exemple réalisé sous CMS, le dialogue du fichier d'appel demande seulement le nom du fichier contenant la matrice de similarité. En version Macintosh, le dialogue demande également le nom du fichier de sortie.

La dernière ligne du groupement (6), qui seule représente l'information à conserver, se lit comme suit. Les 24 stations d'échantillonnage du transect sont représentées par autant de caractères:

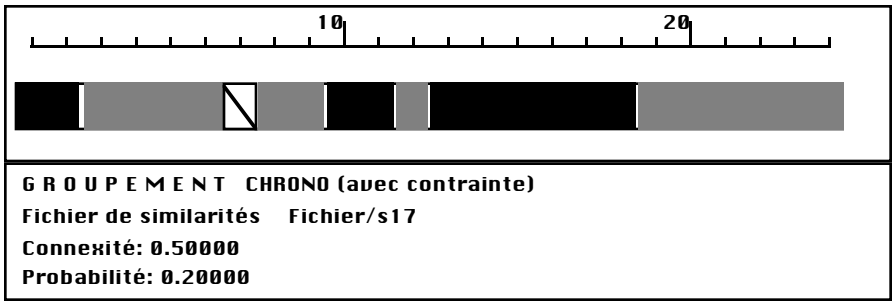
```
AABBBB*BBCC-DDDDDDEEEEEE   S:  0.26667 H: 0.30000
```

La première station se trouve à l'extrémité gauche. Les groupes formés sont représentés par des lettres; ainsi, dans cet exemple, il y a cinq groupes formés, représentés par les lettres A à E. Les stations non groupées sont représentées par des tirets (-) et les singletons par des astérisques (*); la différence réside en ce que les singletons ont été dûment testés par rapport aux groupes situés à leur gauche et à

section “Que fait CHRONO?”, ainsi que ci-dessous). La valeur qui suit “S” représente le niveau de similarité auquel s’est effectuée la dernière fusion, la valeur qui suit “H” représentant la probabilité de l’hypothèse nulle ayant conduit à cette fusion.

La version Macintosh produit l’image suivante qui résume le groupement; le programme permet d’inscrire cette image directement sur un fichier de type PICT, ce qui permet de l’éditer et de l’inclure directement dans une publication. Dans cette image, les groupes successifs sont représentés par des zones alternées de gris et de noir. L’objet 7, représenté par un carré blanc barré, est un singleton; celui-ci se distingue de l’objet 12, qui forme un groupe d’un seul prélèvement et représente un cas spécial. La différence réside dans le fait que l’objet 12 présente des similarités de zéro avec ses voisins immédiats; puisque le groupement s’arrête avant le niveau de similarité $S = 0$, cet objet n’est jamais groupé, et il se retrouve donc seul; comme il n’est pas testé non plus, il n’est donc pas identifié comme singleton. De tels objets peuvent, par leur présence dans une série, interrompre la formation de groupes; lorsqu’il s’en trouve dans une analyse, on doit se demander s’il ne s’agirait pas d’objets aberrants à un titre ou à un autre, auquel cas ils doivent être éliminés de l’étude.

Transect spatial



On pourrait également représenter les objets dans un espace réduit (analyse des correspondances, analyse en coordonnées principales de la matrice S17, cadrage multidimensionnel non-métrique, etc.) et relier par des traits les stations membres d’un même groupe.

```
Quel est le nom du fichier contenant la MATRICE DE SIMILARITES?
(Par defaut: "... data a")
fichier  s17  a

Execution begins...            Annonce le début de l'exécution du programme de tri
Execution begins...     Annonce le début de l'exécution du programme de groupement

 G R O U P E M E N T   C H R O N O L O G I Q U E

 DEPARTEMENT DE SCIENCES BIOLOGIQUES
 UNIVERSITE DE MONTREAL
 C. P. 6128, SUCC "A"
 MONTREAL, QUEBEC  H3C 3J7.

 Reference decrivant la methode:
 Legendre,  P.,  S.   Dallot,  and  L.   Legendre.  1985  --
    Succession of species within a community:  chronological
    clustering,  with applications to marine  and freshwater
    zooplankton.  The American Naturalist, 125 (2): 257-288.

(1)   CONNEXITE DU GROUPEMENT ?
```

```
     0.5
      Connexite:     0.50

(2)   NIVEAU ALPHA POUR LE TEST DE FUSION DES GROUPES ?
     0.20
      Niveau de fusion des groupes (ALPHA):  0.20000

(3)   ELIMINATION DES OBJETS ABERRANTS (O ou N) ?
     n
      Pas d'elimination des objets aberrants.

      LARGEUR DE VOTRE TERMINAL, EN N. DE COLONNES?
     80                          La largeur habituelle d'un écran est de 80 ou 132 caractères

(4,5) DESIREZ-VOUS LES TESTS A POSTERIORI (O ou N) ?
     o

      H  est  la probabilite  que  l'hypothese  principale  soit
      vraie.  Selon celle-ci,  les deux groupes  soumis  au test
      sont un artefact  et devraient  etre fusionnes  en un seul
      groupe. La fusion est accomplie si H est plus eleve que le
      seuil de probabilite ALPHA etabli plus haut par l'usager.

      ------------------AA----   S:  0.84615
      ----------------AABB----   S:  0.84211
      AA--------------BBCC----   S:  0.81818
      AA--------------BBCCC---   S:  0.81481 H: 0.66667
      AABB------------CCDDD---   S:  0.71429
      AABB-----------CCCDDD---   S:  0.66667 H: 0.66667
      AABBB----------CCCDDD---   S:  0.53333 H: 0.66667
      AABBBB---------CCCDDD---   S:  0.53333 H: 1.00000
      AABBBB-------CCDDDEEE---   S:  0.50000
      AABBBB------CCCDDDEEE---   S:  0.50000 H: 0.66667
      AABBBB---CC-DDDEEEFFF---   S:  0.44444
      AABBBB---CC-DDDDDDEEE---   S:  0.42105 H: 0.40000
      AABBBB-CCDD-EEEEEEFFF---   S:  0.40000
      AABBBB-CCCC-DDDDDDEEE---   S:  0.30769 H: 0.33333
      AABBBB-CCCC-DDDDDDEEEFF-   S:  0.30000
      AABBBB-CCCC-DDDDDDEEEFFF   S:  0.28571 H: 0.66667
      AABBBB-CCCC-DDDDDDEEEEEE   S:  0.26667 H: 0.30000
      L'OBJET:  7 EST ELIMINE  H: 0.20000 0.20000

      ------*-----------AA----   S:  0.84615
      ------*---------AABB----   S:  0.84211
      AA----*---------BBCC----   S:  0.81818
      AA----*---------BBCCC---   S:  0.81481 H: 0.66667
      AABB--*---------CCDDD---   S:  0.71429
      AABB--*--------CCCDDD---   S:  0.66667 H: 0.66667
      AABBB-*--------CCCDDD---   S:  0.53333 H: 0.66667
      AABBBB*--------CCCDDD---   S:  0.53333 H: 1.00000
      AABBBB*------CCDDDEEE---   S:  0.50000
      AABBBB*-----CCCDDDEEE---   S:  0.50000 H: 0.66667
      AABBBB*--CC-DDDEEEFFF---   S:  0.44444
      AABBBB*--CC-DDDDDDEEE---   S:  0.42105 H: 0.40000
```

```
     AABBBB*BBCC-DDDDDDEEE---   S:  0.40000 H: 0.26667
     AABBBB*BBCC-DDDDDDEEEFF-   S:  0.30000
     AABBBB*BBCC-DDDDDDEEEFFF   S:  0.28571 H: 0.66667
(6)  AABBBB*BBCC-DDDDDDEEEEEE   S:  0.26667 H: 0.30000
```

Résultat du groupement

```
     TEMPS ECOULE: 0.7143 SEC

(4)  EXPANSION DES GROUPES

          [   1 ..    2]
          [   1 ..    3]  H:    0.66667
          [etc.]
          [   1 ..    9]  H:    1.00000
          [   1 ..   10]  H:    0.44444
          [   1 ..   11]  H:    0.30000

     etc.
     etc.

          [  19 ..   24]
          [  18 ..   24]  H:    0.85714
          [  17 ..   24]  H:    1.00000
          [  16 ..   24]  H:    1.00000

     TEMPS ECOULE: 1.0083 SEC

(5)  TESTS ENTRE LES GROUPES
     [  1 ..   2] contre  [  3 ..   9]  H: 0.03571
                          [ 10 ..  11]  H: 0.33333
                          [ 12 ..  12]  H: 0.33333
                          [ 13 ..  18]  H: 0.03571
                          [ 19 ..  24]  H: 0.03571
     [  3 ..   9] contre  [ 10 ..  11]  H: 0.03571
                          [ 12 ..  12]  H: 0.14286
                          [ 13 ..  18]  H: 0.14286
                          [ 19 ..  24]  H: 0.02814
     [ 10 ..  11] contre  [ 12 ..  12]  H: 0.33333
                          [ 13 ..  18]  H: 0.10714
                          [ 19 ..  24]  H: 0.03571
     [ 12 ..  12] contre  [ 13 ..  18]  H: 0.14286
                          [ 19 ..  24]  H: 0.14286
     [ 13 ..  18] contre  [ 19 ..  24]  H: 0.07359
```

Le premier groupe [1 .. 2] sert de point de départ à l'expansion

Expansion réalisée de 1 à 11

À tour de rôle, chaque groupe formé sert de point de départ à l'expansion

Le dernier groupe [19 .. 24] sert de point de départ à l'expansion

Expansion réalisée de 16 à 24

Pas de fusion car H ≤ alpha
*
**
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
**
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha
Pas de fusion car H ≤ alpha

* *Cette valeur représente la plus faible probabilité de fusion possible entre ces deux groupes, attendu leur taille (voir le tableau 1). Elle ne représente donc pas nécessairement le non-rejet de* H_0.

** *Il s'agit également de la plus faible valeur possible de*

```
TESTS SUR LES OBJETS ELIMINES

[  7] contre  [  1 ..   2]  H: 0.66667                                  *
              [  3 ..   9]  H: 0.14286          Pas de fusion car H ≤ alpha
              [ 10 ..  11]  H: 0.66667                                  *
              [ 13 ..  18]  H: 0.14286          Pas de fusion car H ≤ alpha
              [ 19 ..  24]  H: 0.28571     ⇐ Fusion de [7] et de [19 .. 24]

TEMPS ECOULE: 1.6521 SEC


FICHIER D'ENTREE:

    NOMBRE D'OBJETS :   24                  Identification du fichier d'entrée
    NOMBRE DE VARIABLES :   41
    TITRE :  Fichier de donnees
    DATE : 02/04/91
    FONCTION :  s17

Fin du programme.
```

COCOPAN

Que fait COCOPAN ?

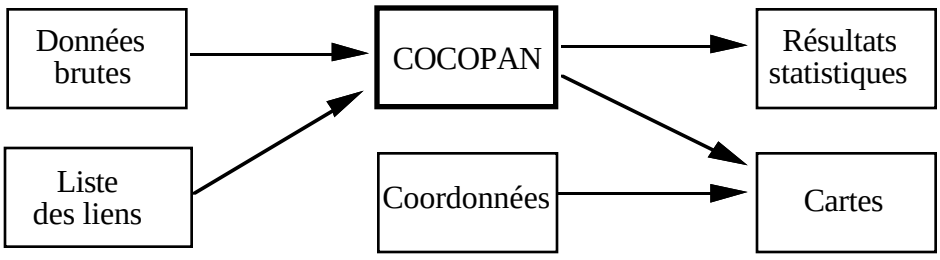
Le programme COCOPAN permet de réaliser une analyse de variance à un critère de classification pour des données quantitatives spatialement autocorrélées, lorsque le critère de classification consiste en une partition du territoire à l'étude en des sous-régions connexes — par exemple des pays, des comtés, des groupes linguistiques, des subdivisions géomorphologiques, et ainsi de suite, comme on en rencontre dans nombre de problèmes dont les données peuvent être représentées sur une carte. La méthode a été décrite par Legendre, Oden, Sokal, Vaudor et Kim (1990). L'acronyme COCOPAN vient du nom anglais de la méthode, ***Co****ntiguity-****co****nstrained* ***p****ermutational* ***A*** *NOVA*.

Le principe de ce test par permutations consiste à garder les localités immobiles, chacune conservant ses valeurs des différentes variables, de façon à préserver la structure d'autocorrélation. On permute plutôt le critère de classification, soit la division de la carte en sous-régions, avec les contraintes suivantes: chaque pseudo-région doit contenir le même nombre de localités que la région d'origine qu'elle représente; chaque pseudo-région doit demeurer connexe, i.e., former une surface continue sur la pseudo-carte; enfin, les pseudo-régions doivent occuper toute la carte d'origine, sans omission de localités ni dépassement. Le programme contient deux algorithmes permettant de résoudre ce problème informatique: l'algorithme des cercles concentriques, conçu par Alain Vaudor, et la méthode de l'arborescence aléatoire, développée par Junhyong Kim.

Plusieurs variables peuvent être analysées en une seule passe. La statistique utilisée dans le test permutationnel est la somme, pour tous les groupes, des sommes de carrés intragroupes (SCE). Après chaque permutation, on recalcule la statistique SCE pour cette pseudo-carte; on compare enfin la valeur SCE obtenue pour la vraie carte à la distribution des valeurs de SCE obtenues pour les pseudo-cartes. Le test est donc unilatéral et la région critique se trouve à l'extrémité gauche de la distribution.

Si vous utilisez une version du programme pour grands ordinateurs, vérifiez les constantes au début du programme (déclaration CONST) pour vous assurer qu'il pourra traiter votre problème; vérifiez en particulier la valeur de MAXLOC (nombre maximum de localités), MAXGROUPES (nombre maximum de groupes, limité à 255 dans la version Macintosh) et MAXVAR (nombre maximum de variables). Vous pouvez changer ces valeurs pour traiter des problèmes plus importants. Choisissez également la langue de conversation du programme: LANG = 1 pour le français.

Fichiers d'entrée et de sortie



Outre les fichiers INPUT et OUTPUT qui représentent le clavier et l'écran du terminal ou du microordinateur, trois fichiers d'entrée sont nécessaires à ce programme; celui-ci produit, par ailleurs, deux fichiers de sortie. Le premier fichier d'entrée est le même qu'en analyse de variance ordinaire, soit les différentes variables à analyser ainsi que le critère de classification. Pour tenir compte de la structure spatiale, un second fichier est nécessaire, qui indique au programme quelles sont les localités

voisines sur la carte. Enfin, si on désire obtenir des cartes, il faut fournir au programme un troisième fichier précisant les coordonnées géographiques de chaque localité. En sortie, on peut obtenir un fichier de statistiques ainsi que des cartes, qui sont écrites sur un second fichier dans les versions pour grands ordinateurs. Tous ces fichiers sont écrits en caractères lisibles (ASCII).

(1) Fichier des données

Les lignes de ce fichier correspondent aux différentes localités (objets). Les premières *N* colonnes sont les *N* variables à analyser; la dernière colonne contient le critère de classification (groupe géographique), codé en entiers de 1 jusqu'au nombre de groupes *k*; cette valeur doit être strictement inférieure à la constante MAXGROUPES, dans la liste des constantes en tête du programme, pour les versions VMS/CMS. Ce fichier, qui porte le nom de DATAFILE dans le programme PASCAL, devrait être compatible avec les formats d'entrée de la plupart des logiciels statistiques standards, ce qui permet de réaliser aisément une ANOVA standard, pour fins de comparaison. Le programme COCOPAN ne peut traiter les données manquantes; l'utilisateur doit s'assurer que les localités avec données manquantes ont été éliminées des trois fichiers d'entrée, ou que les valeurs manquantes ont été estimées, par interpolation ou autre méthode, avant cette analyse.

(2) Liste des liens entre les objets

Ce fichier, qui porte le nom de LINKS dans le programme PASCAL, fournit au programme une liste des liens entre paires de localités voisines. Chaque lien est représenté par une paire de numéros de localités, écrits en format libre et séparés par au moins un espace. Ce fichier peut être fabriqué à l'aide du programme CONNEXIONS (version Macintosh) ou du programme AUTOCOR (versions VMS ou CMS); voir la description de ces programmes. Puisqu'il est écrit en ASCII, ce fichier peut être édité par l'usager (addition ou élimination de certains liens), ou encore écrit entièrement par lui, à l'aide de son éditeur ASCII. Cet élément de flexibilité permet de traiter des problèmes représentant un volume plutôt qu'une surface, pour autant que l'on fournisse au programme une liste de liens représentant les relations de voisinage entre points-objets en trois dimensions.

(3) Liste des coordonnées (X, Y)

Ce fichier, qui porte le nom de COORD dans le programme PASCAL, contient la liste des coordonnées géographiques (X et Y) des localités. Il est requis si l'on désire imprimer des cartes, soit la carte d'origine ainsi que les cartes permutées, ainsi que pour le calcul de la statistique de Diamètre de l'Ensemble (DE) pour chaque pseudo-groupe. Pour que les cartes soient imprimées correctement, les coordonnées en abscisse doivent aller de la droite vers la gauche, comme les longitudes à l'ouest de Greenwich, et les valeurs en ordonnée du bas vers le haut, comme les latitudes de l'hémisphère nord. Autrement, les cartes pourront être inversées. La version Macintosh peut tourner les cartes, si nécessaire, pour les adapter à la forme de l'écran.

(4) Fichier des résultats statistiques

Le premier fichier de résultats, qui porte le nom de STATIS dans le programme PASCAL, contient les statistiques détaillées (voir ci-dessous).

(5) Fichier des cartes

Ce fichier, qui porte le nom de GRAPHICS dans le programme PASCAL, est optionnel et ne sera fourni que si l'usager demande que les cartes soient produites. Il s'agit d'un fichier séparé dans les versions pour grands ordinateurs; dans la version Macintosh, les cartes sont produites directement à l'écran. On peut ainsi reproduire et examiner la carte d'origine ainsi que les cartes permutées (pseudo-cartes). Voir l'exemple ci-dessous.

Les questions du programme

Les questions présentées par le programme sont décrites dans les paragraphes qui suivent.

(1) “Combien de permutations de la carte faut-il réaliser ?” — On indique combien de permutations sont requises. Puisque les probabilités sont calculées en incluant les statistiques de la vraie carte parmi les résultats que l’on peut obtenir sous H_0 (méthode de Hope, 1968: voir plus bas), il convient de compter une permutation de moins que le nombre de cas désirés. Par exemple, pour un total de 1000 statistiques dans la distribution, il faut demander 999 permutations aléatoires des régions définies par le critère de classification. Selon la taille du problème, la vitesse de l’ordinateur disponible et la précision désirée, on demande souvent de 249 à 999 permutations, ou plus.

(2) “Initialisation du générateur de nombres aléatoires: Tapez un ENTIER entre 1 et 100.” — Avant que ne commencent les calculs proprement dits, le générateur de nombres pseudo-aléatoires est appelé “pour rien” le nombre de fois spécifié en réponse à cette question; chaque appel change le point de démarrage du générateur pour la valeur suivante. Le changement du point de démarrage de ce générateur est un moyen simple d’obtenir un jeu différent de cartes permutées, pour un même problème.

(3) “Méthode des cercles concentriques, plutôt que l'arborescence aléatoire ? (Tapez O pour la méthode des cercles concentriques, N pour l'arborescence aléatoire.)” — Les deux algorithmes sont décrits en annexe de l’article cité plus haut, décrivant la méthode COCOPAN. Leurs avantages respectifs sont les suivants: la méthode des cercles concentriques produit des groupes moins allongés, alors que l’algorithme de l’arborescence aléatoire est plus rapide; il a été employé pour traiter des problèmes comportant plus de 1300 localités, et ceci est encore loin de sa limite.

(4) Si on a choisi l’algorithme des cercles concentriques, le programme présente la question suivante: “Combien de fois peut-on tenter de déplacer les groupes ? (Recommandé: 3-5)” — Ceci correspond à l’étape 1.3.2 dans la description de l’algorithme, en annexe de l’article (*op. cit.*). Après avoir essayé le nombre de fois spécifié en réponse à cette question, la carte est mise de côté si elle n’a pas été complétée.

(5) Également pour l’algorithme des cercles concentriques: “Jusqu'à combien de points manquants doit-on tenter de compléter une carte de force ? (Recommandé: 1)” — Il s’agit ici de l’étape 1.3.3 dans la description de l’algorithme. S’il ne manque que quelques points pour compléter la dernière pseudo-région (le nombre maximum de points manquants étant spécifié en réponse à cette question), le programme trouve le chemin le plus court entre les cases encore inoccupées et le pseudo-groupe à compléter, et déplace toute la chaîne des assignations le long de ce parcours pour forcer la création de cases libres adjacentes à ce groupe, lui permettant ainsi de compléter son jeu de localités.

(6) “FICHIER DES CARTES: Dessiner les cartes ? (O ou N)” — L’usager peut demander que les cartes (la carte d’origine ainsi que toutes les pseudo-cartes) soient copiées dans un fichier (versions VMS ou CMS), ou encore, dans la version Macintosh, qu’elles apparaissent à l’écran, d’où elles pourront être imprimées.

(7) “Inscrire en représentation codée quelles localités forment chaque groupe ? (O ou N)” — Cette sortie est écrite dans le fichier des cartes (versions VMS et CMS) ou dans le fichier des résultats statistiques (version Macintosh). Elle se présente comme suit:

```
----++++--++++++-+++++----+-----------------------------------   16
---------------------------++++----++++----++++----++++----+++--+++   19
++++----++------+-----++++---++++---++++---++++----++++---++---   29
-------------------------------------------------------------   0
```

Dans cet exemple, il y a trois groupes correspondant respectivement à la première, la deuxième et la troisième ligne, identifiées par les cardinalités 16, 19 et 29. La quatrième ligne (cardinalité de 0) représenterait les localités non assignées, dans le cas de cartes non complétées. Pour chaque groupe de la vraie carte (et de la même façon pour chaque groupe de chacune des pseudo-cartes), on trouve une ligne de caractères représentant toutes les localités de l'étude, dans l'ordre où elles apparaissent dans le fichier de données. Ces caractères sont des plus (+) pour les localités présentes dans le groupe en question, et des moins (-) pour les localités ne faisant pas partie de ce groupe. Lorsqu'il y a plus de 80 localités dans l'étude, la chaîne de caractères représentant chaque groupe prend plus d'une ligne; dans ce cas, la première ligne est d'abord écrite pour tous les groupes, puis la seconde ligne pour tous les groupes, et ainsi de suite. Puisque cette représentation codée contient toute l'information quant à l'appartenance des localités, d'abord pour la vraie carte puis pour chacune des pseudo-cartes, elle peut donc être employée comme fichier d'entrée pour d'autres programmes. Ainsi, les pseudo-cartes produites par COCOPAN pourront être employées pour calculer des statistiques additionnelles sur la forme des pseudo-groupes ou la distribution des localités, ou encore pour réaliser d'autres analyses basées sur les pseudo-cartes COCOPAN (MANOVA, analyse de variance non paramétrique, analyse discriminante, etc.). Cette sortie n'est pas utile dans les applications routinières de COCOPAN.

(8) "Inscrire quel no de groupe a été attribué à chaque localité ? (O ou N)" — Voici une autre sortie qui contient toute l'information quant à la position des localités parmi les groupes, pour la vraie carte ainsi que pour chacune des pseudo-cartes. Cette sortie est écrite dans le fichier des cartes (versions VMS ou CMS) ou encore dans le fichier des résultats statistiques (version Macintosh). Elle se présente comme une liste des localités, avec le numéro du groupe attribué à chacune d'elles. En voici un exemple:

```
3  3  3  3  1  1  1  1  3  3  1  1  1  1  1  1
3  1  1  1  1  1  3  3  3  3  1  2  2  2  3  3
3  3  2  2  2  3  3  3  3  2  2  2  3  3  3  3
2  2  2  2  3  3  3  3  2  2  2  3  3  2  2  2
```

Dans cet exemple, on apprend que les quatre premières localités appartiennent au groupe 3, les quatre suivantes au groupe 1, et ainsi de suite (comparer avec la représentation codée du paragraphe précédent). Cette sortie peut servir aux mêmes fins que la sortie du paragraphe précédent. Elle n'est pas utile dans les applications routinières de COCOPAN.

(9) "DEBUG: Inscrire toutes les cartes et/ou les listes de bits, même pour les cartes rejetées ? (O ou N)" — Il s'agit ici des sorties décrites aux points (6) à (8) ci-dessus; cette question n'apparaît que si on a choisi l'algorithme des cercles concentriques. On répond 'Oui' si on désire en savoir davantage sur les cartes qui ont été rejetées. Elle a été prévue en vue d'identifier les problèmes qui empêchent la formation de pseudo-cartes, tels que les étranglements dans la chaîne des liens de voisinage.

(10) "Largeur des cartes (en nombre de caractères) ?" — Dans les versions VMS et CMS du programme, vous pouvez choisir la largeur des cartes requises à la question 6, en fonction du terminal que vous utilisez et de la taille des cartes désirées. Ces cartes simples sont composées à l'aide des caractères du clavier et leur largeur est calculée en nombre de caractères (voir l'exemple plus bas). Cette question n'apparaît pas dans la version Macintosh, qui produit à l'écran des cartes au trait pouvant être reproduites sur imprimante Laser.

(11) "Nombre de variables à analyser ?" — Indiquer ici combien il y a de variables dans le fichier de données, **sans compter** le critère de classification qui forme la dernière colonne de ce fichier (voir la description du fichier des données, ci-dessus).

(12) "Diamètre des ensembles (*Set Diameter*) ? (O ou N)" — Cette statistique (DE) est décrite plus en détail dans l'article (*op. cit.*, voir *Set Diameter*). Il s'agit du diamètre du plus petit cercle qui contient toutes les localités membres d'un groupe ou d'un pseudo-groupe. Ce diamètre est calculé en tenant compte de la courbure de la terre, sous l'hypothèse que les coordonnées X et Y du fichier de coordonnées sont exprimées en degrés; le diamètre, pour sa part, est exprimé en minutes d'arcs (ce qui

est équivalent à des milles marins). Ces statistiques permettent de comparer le diamètre des pseudo-groupes produits par le programme au diamètre des groupes d'origine.

(13) "Probabilités des diamètres d'ensembles ? (O ou N)" — Si on choisit cette option, le fichier des statistiques contiendra un tableau donnant la probabilité, pour chaque groupe, de trouver parmi les permutations effectuées des pseudo-groupes ayant un diamètre plus petit ou égal au diamètre de ce groupe sur la vraie carte. Voir les remarques ci-dessous sur le calcul des probabilités permutationnelles; voir aussi l'exemple.

(14) "Diamètre en nombre de liens (*Path Length*) ? (O ou N) (Attention: temps de calcul élevé s'il y a beaucoup de points.)" — Cette seconde statistique (DNL) de la forme des groupes n'est pas décrite dans l'article. Le but est le même qu'avec la mesure des diamètres d'ensembles: il s'agit de comparer le diamètre des pseudo-groupes produits par le programme au diamètre des groupes d'origine. Le diamètre est cependant mesuré différemment; la mesure est le nombre minimum de liens (voir la description du fichier des liens ci-dessus) nécessaires pour rejoindre les deux localités les plus éloignées (également en terme de nombre de liens) dans le groupe ou le pseudo-groupe. Le message rappelle à l'usager que cette statistique est très coûteuse à calculer pour de grands jeux de données.

(15) "Probabilités des diamètres en nombre de liens ? (O ou N)" — Ces probabilités sont calculées comme décrit en (13), pour les diamètres en nombre de liens.

(16) "Statistiques pour CHAQUE carte ? (O ou N)" — Si on choisit cette option, la statistique de Somme des Carrés des Écarts (SCE) est rapportée pour chaque groupe (i) séparément [SCE(i)] et pour l'ensemble du problème [SCE = somme des SCE(i)], et ce pour chaque carte (c'est-à-dire pour la vraie carte et pour chacune des pseudo-cartes). Si elles ont été demandées aux questions (12) et (14), les statistiques de forme des groupes, DE et DNL, sont aussi fournies pour chaque carte. Voir l'exemple ci-dessous.

Que l'on ait choisi ou non cette option, un tableau-synthèse est présenté à l'écran **ainsi que** dans le fichier des statistiques. Voir l'exemple ci-dessous. Pour chaque groupe, ce tableau présente la probabilité de trouver, parmi les permutations effectuées, des pseudo-groupes ayant une SCE(i) plus petite ou égale à celle de ce même groupe sur la vraie carte. Ceci nous informe sur l'homogénéité interne de chaque groupe de la vraie carte, par comparaison à l'homogénéité de groupes connexes, formés au hasard sur la carte, possédant le même nombre de localités.

La dernière ligne de ce tableau (TOTAL) présente les résultats principaux de l'analyse de variance, c'est-à-dire la probabilité de trouver parmi les permutations effectuées des valeurs SCE plus petites ou égales à celle de la vraie carte. Ce tableau est répété pour chaque variable de l'étude.

(17) "Statistiques quant à la fréquence des localités dans chaque groupe? (O ou N)" — Pour chaque groupe à tour de rôle, une liste est écrite sur le fichier des statistiques qui nous informe sur le nombre de fois où chaque localité a été choisie pour faire partie du groupe en question; par exemple (pour 500 permutations aléatoires):

```
Fréquence des localités dans le groupe 1
 100  107  101  112  120  124  119  117  107  120  126  136
 139  135  119  115  113  135  141  136  135  127  114  138
 159  148  139  113   97   87  111  157  149  135  125  128
 109  117  148  144  139  124  119  105  122  147  141  136
 130  119  111   94  123  134  145  151  128  127   98  109
 112  127  129  128
Fréquence des localités dans le groupe 2
 170  166  167  149  130  129  131  132  169  167  147  163
 136  128  133  133  159  148  140  148  134  128  166  164
```

```
 134  128  139  133  127  124  176  146  135  123  144  140
 129  172  167  129  125  138  139  143  182  170  163  136
 137  144  148  147  176  176  158  144  154  148  150  185
 187  162  157  148
Fréquence des localités dans le groupe 3
 230  227  232  239  250  247  250  251  224  213  227  201
 225  237  248  252  228  217  219  216  231  245  220  198
 207  224  222  254  276  289  213  197  216  242  231  232
 262  211  185  227  236  238  242  252  196  183  196  228
 233  237  241  259  201  190  197  205  218  225  252  206
 201  211  214  224
```

Pour ce problème comportant 64 localités (les trois groupes comportant respectivement 16, 19 et 29 localités), on apprend par exemple que la première localité de la liste a été choisie 100 fois (sur 500 tentatives) pour faire partie du groupe 1, alors qu'à 170 reprises elle a fait partie du groupe 2 et à 230 reprises du groupe 3. Pour les problèmes où la densité des connexions n'est pas uniforme, cette liste informe le chercheur si l'attribution des localités aux différents groupes s'est faite au hasard ou non. Ce problème est discuté plus à fond dans la section 3.2 de l'article (*op. cit.*), où il est montré que l'attribution des localités aux pseudo-groupes peut être inégale dans des réseaux à densité de connexion très variable.

(18) Si on a choisi l'algorithme de l'arborescence aléatoire, la question suivante est posée: "Combien d'arbres aléatoires permettez-vous d'avorter avant que le programme ne s'arrête? On recommande 10*(N de permutations)." — Il peut être nécessaire dans certains problèmes d'accroître cette valeur pour permettre de compléter les pseudo-cartes. Il y a cependant peu de chances que cela se produise. Prière de nous rapporter ces cas.

Les probabilités obtenues par permutations sont calculées selon la méthode de Hope (1968), méthode recommandée également par Edgington (1987); celle-ci consiste à inclure la valeur observée parmi les "Égaux" de la distribution de référence, de sorte qu'il n'est jamais possible d'obtenir 0% de valeurs "plus petites ou égales" à la valeur observée. Selon Edgington, cette façon de faire introduit un biais mais elle a le mérite d'être valide. La précision de cette probabilité est l'inverse du nombre de permutations demandées par l'usager.

Exemple

L'exemple ci-dessous illustre l'utilisation du programme sur grands ordinateurs (système VMS ou CMS; cet exemple a été réalisé sous VMS). Le programme de lancement demande d'abord à l'usager d'identifier les fichiers qui seront utilisés; les réponses sont soulignées. Puis, après l'en-tête du programme, viennent les questions posées par le programme lui-même pour identifier quelles sont les options de calcul que désire l'usager.

```
Programme COCOPAN

Quel est le nom du fichier principal de DONNEES, contenant
les variables ainsi que le critere de classification?
(Par defaut: "... data a")
donnees

Quel est le nom du fichier contenant les LIENS entre localites?
(Par defaut: "... data a")
liens

Quel est le nom du fichier des COORDONNEES geographiques?
```

```
Ce fichier n'est requis que si vous desirez imprimer les cartes,
ou encore si vous demandez a calculer le diametre des groupes
(defaults are "... data a")
coordxy

Sur quel fichier les CARTES devront-elles etre imprimees?
(Optionnel; par defaut: "CARTES data a")


Sur quel fichier les STATISTIQUES detaillees devront-elles
etre inscrites? (Par defaut: "STATIS data a")


P r o g r a m m e  C O C O P A N -- C a r t e s

(ANOVA par permutations sous contrainte de contiguite)

Reference:
Legendre, P., N.L. Oden, R.R. Sokal, A. Vaudor and J. Kim. 1990.
   Approximate analysis of variance of spatially autocorrelated
   regional data. J. Class. 7: 53-75.

Credits --
Programme et algorithme des cercles concentriques: Alain Vaudor,
Departement de sciences biologiques,
Universite de Montreal,
C.P. 6128, Succursale A,
Montreal, Quebec H3C 3J7.

Algorithme de l'arborescence aleatoire: Junhyong Kim,
State University of New York at Stony Brook.

Combien de permutations de la carte faut-il realiser ?
999
Initialisation du generateur de nombres aleatoires:
Tapez un ENTIER entre 1 et 100.
50
Methode des cercles concentriques, plutot que l'arborescence aleatoire ?
(Tapez O pour la methode des cercles concentriques,
       N pour l'arborescence aleatoire.)
n
FICHIER DES CARTES:
   Dessiner les cartes ? (O ou N)
n
   Inscrire en representation codee quelles localites forment chaque groupe?
   (O ou N)
n
   Inscrire quel no de groupe a ete attribue a chaque localite ? (O ou N)
n
Nombre de variables a analyser ?
1
FICHIER DES STATISTIQUES:
   Diametre des ensembles (Set Diameter) ? (O ou N)
o
```

```
   Probabilites des diametres d'ensembles ? (O ou N)
o
   Diametre en nombre de liens (Path Length) ? (O ou N)
   (Attention: temps de calcul eleve s'il y a beaucoup de points.)
o
   Probabilites des diametres en nombre de liens ? (O ou N)
o
   Statistiques pour CHAQUE carte ? (O ou N)
o
   Statistiques quant a la frequence des localites dans chaque groupe?
   (O ou N)
n
Combien d'arbres aleatoires permettez-vous d'avorter avant que le
programme ne s'arrete? On recommande 10*(N de permutations).
10000

Probabilites des statistiques SCE:

 Variable  Groupe   Plus petits  Egaux  N.cartes Prob(H0)
     1         A       597         1      1000    0.5980
     1         B       170         1      1000    0.1710
     1         C       249         1      1000    0.2500
     1     total        22         1      1000    0.0230
Fin du programme.
```

Contenu du fichier de résultats statistiques

Dans ce fichier de résultats, les numéros à l'extrême gauche réfèrent aux numéros de la section "Les questions du programme" ci-dessus.

```
C O C O P A N - Fichier des statistiques
Le fichier des liens contient:
   Nombre total de liens:       203
   Nombre moyen de liens:   3.17187
   Ecart type           :   1.84493
   Variance du n. liens :   3.40377

DE  est le diametre de l'ensemble (pour chaque groupe, en min. d'arc)
DNL est le diametre en nombre de liens (pour chaque groupe)
SCE pour chaque groupe, puis pour l'ANOVA, pour les variables V(1) a V(n)

N. loc.          16        19        29 [Nombre de localités dans chaque groupe]

(16) Carte no    0                                    [La carte "0" est la vraie carte]

 DE         402.259   423.800   509.695
 DNL              6         6         9
 SCE v 1    117.726    94.234   188.758   400.719 ["v 1" signifie "variable 1"]

Carte no    1

 DE         515.492   515.492   515.492
 DNL              8         8         7
```

```
     Carte no    2

      DE         423.219   383.449   592.866
      DNL              5         5         9
      SCE v 1    121.579    55.596   236.485   413.660

     [etc.]

     Carte no  998

      DE         383.569   483.294   636.466
      DNL              5         7        11
      SCE v 1     42.540   214.898   245.102   502.540
     Carte no  999

      DE         423.800   383.449   515.492
      DNL              5         5         8
      SCE v 1    102.966    58.035   322.877   483.878

(16) Probabilites des statistiques de SCE:
```

[Les “Égaux” incluent la carte “0”, qui est la vraie carte]

```
      Variable  Groupe   Plus petits  Egaux  N.cartes Prob(H0)
          1         A       597         1      1000    0.5980
          1         B       170         1      1000    0.1710
          1         C       249         1      1000    0.2500
          1     total        22         1      1000    0.0230

(13) Probabilites des statistiques DE:

          Groupe  Plus petits  Egaux Prob(H0)
             1       462        23    0.4850
             2       257        43    0.3000
             3       137        13    0.1500

(15) Probabilites des statistiques DNL:

          Groupe  Plus petits  Egaux Prob(H0)
             1       389       216    0.6050
             2       195       203    0.3980
             3       582       172    0.7540
```

Fichier des cartes

Les trois cartes ci-dessous illustrent le type de cartes qui peuvent être imprimées sur une imprimante régulière, à l'aide des versions VMS et CMS du programme. Ces cartes montrent comment les groupes d'origine (carte 0) peuvent être déplacés par l'algorithme. La version Macintosh, quant à elle, produit des cartes au trait dans lesquelles chaque groupe est délimité par une enveloppe.

```
Carte no   0              Carte no   1              Carte no   2
-----------------------   -----------------------   -----------------------
!C  C  B  B  B         !   !B  B  A  A  A         !   !C  C  C  C  C         !
!                      !   !                      !   !                      !
!C  C  C  C  B  B  B   !   !B  B  B  A  A  A  A   !   !C  C  C  C  C  C  C   !
!                      !   !                      !   !                      !
!C  C  C  C  B  B  B  B!   !B  B  B  A  A  A  A  A!   !C  C  C  C  C  C  C  C!
!                      !   !                      !   !                      !
!C  C  C  C  B  B  B   !   !B  B  B  B  A  A  A   !   !B  C  C  C  C  C  C   !
!                      !   !                      !   !                      !
!C  C  C  C  B  B  B   !   !B  B  B  B  A  C  C   !   !B  B  A  A  A  A  C   !
!                      !   !                      !   !                      !
!C  C  C  C  A  B  B  B!   !B  B  C  C  C  C  C  C!   !B  B  B  A  A  A  C  C!
!                      !   !                      !   !                      !
!C  A  A  A  A  A      !   !B  C  C  C  C  C      !   !B  B  A  A  A  A      !
!                      !   !                      !   !                      !
!C  C  A  A  A  A  A  A!   !C  C  C  C  C  C  C  C!   !B  B  B  B  B  A  A  A!
!                      !   !                      !   !                      !
!C  C  C  C  A  A  A  A!   !C  C  C  C  C  C  C  C!   !B  B  B  B  B  B  A  A!
-----------------------   -----------------------   -----------------------
```

CONNEXIONS Macintosh

Que fait CONNEXIONS ?

Ce programme permet de réaliser différents schémas de connexions entre localités voisines dans l'espace (1 ou 2 dimensions) et d'inscrire les liens de proximités dans un fichier; certains programmes d'analyse spatiale, tels que le programme d'autocorrélation spatiale, les programmes de groupement avec contrainte de contiguïté spatiale BIOGÉO et K-MEANS, ainsi que la méthode d'analyse de variance COCOPAN, utilisent ces fichiers comme information relative aux liens de voisinage qui existent entre les localités. Le programme CONNEXIONS n'existe qu'en version Macintosh. La plupart de ses fonctions sont disponibles, pour les versions CMS et VMS, dans le programme AUTOCOR.

Lorsque les points forment une grille régulière sur la carte, il est aisé de relier les plus proches voisins par des schémas de connexions simples nommés par référence au jeu d'échecs (Cliff & Ord, 1981): mouvement de la tour (en carré), du fou (en diagonale) ou du roi (appelé aussi mouvement de la reine: en carré et en diagonale).

Lorsque les localités sont disposées de façon irrégulière, on peut employer des méthodes de connexion géométriques telles que le critère de connexion de Gabriel (Gabriel & Sokal, 1969), la triangulation de Delaunay (Upton & Fingleton, 1985) ou le schéma de voisinage relatif. Il existe une relation d'inclusion entre ces schémas de connexion: tous les liens qui peuvent être établis par le schéma de voisinage relatif sont aussi inclus dans le schéma de connexion de Gabriel, ceux-ci se retrouvant tous dans la triangulation de Delaunay:

Voisinage relatif ⊃ Critère de Gabriel ⊃ Triangulation de Delaunay

Fichiers d'entrée et de sortie

(1) Fichier des coordonnées

Pour la triangulation de Delaunay, le schéma de connexion de Gabriel et le schéma de voisinage relatif, il faut fournir au programme un fichier des coordonnées géographiques des localités. Chaque ligne de ce fichier doit contenir deux informations, comme suit:

Coordonnée en X Coordonnée en Y

Les coordonnées doivent être fournies sous la forme de nombres entiers ou de réels (*i.e.*, nombres décimaux) et non sous la forme de degrés-minutes-secondes. Ces données sont lues en format libre; autrement dit, le nombre d'espaces avant ou après chaque chiffre n'importe pas. Pour une grille régulière de localités, aucun fichier d'entrée n'est requis.

(2) Sortie: Fichier des liens

Ce fichier ASCII contient une liste de liens entre paires d'objets (points) voisins, tel que le permet la connexion (option) qui a été utilisée lors de l'exécution du programme. Chaque lien est représenté par le numéro des deux points qu'il relie. L'exemple qui suit correspond à une grille régulière de 4 lignes et 5 colonnes (20 localités), mouvement du roi:

```
 1   2     2   3     3   4     4   5     6   7     7   8     8   9     9  10
11  12    12  13    13  14    14  15    16  17    17  18    18  19    19  20
 6   1     7   2     8   3     9   4    10   5    11   6    12   7    13   8
14   9    15  10    16  11    17  12    18  13    19  14    20  15     6   2
 7   3     8   4     9   5    11   7    12   8    13   9    14  10    16  12
17  13    18  14    19  15     7   1     8   2     9   3    10   4    12   6
13   7    14   8    15   9    17  11    18  12    19  13    20  14
```

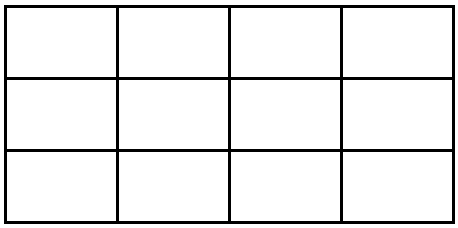
Il est à noter que l'utilisateur peut éditer ce fichier ASCII; il peut enlever des liens ou en ajouter, selon les besoins de son étude.

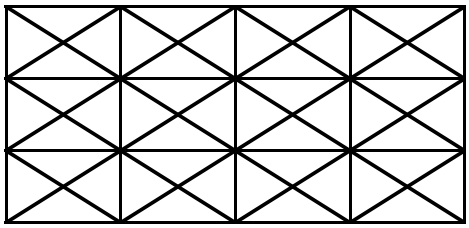
Options

L'exposé des options du programme contient également des exemples d'utilisation.

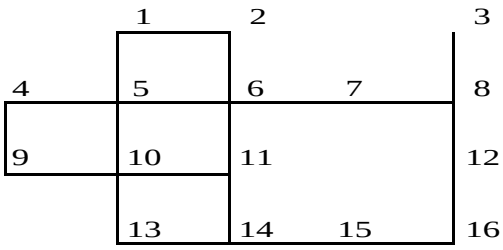
(1) Grille régulière

Si les points sont disposés selon une grille rectangulaire régulière, il n'est pas nécessaire de fournir une liste des coordonnées. Le programme demande d'abord la taille de cette grille (nombre de colonnes, nombre de lignes), puis les traits qui sont désirés: traits horizontaux, verticaux, de pente positive et de pente négative. La figure de gauche, ci-dessous, illustre une connexion de la tour (traits horizontaux et verticaux), alors que celle de droite présente une connexion du roi (traits dans les quatre directions). Pour l'exemple de droite, la liste des liens est fournie ci-dessus, les points étant numérotés par ligne, de 1 à 20, comme on lit un texte.





Étant donné une grille régulière d'une certaine taille, le programme offre la possibilité d'éliminer certains points de la grille. On doit d'abord indiquer combien de points devront être éliminés, puis identifier ces points, en supposant une numérotation de gauche à droite sur chaque ligne, et du haut vers le bas. Ainsi, on pourrait éliminer les points nos 1, 4, 14 et 16 des grilles de 20 points ci-dessus et obtenir le schéma suivant (connexion de la tour) qui ne contient plus que 16 points:

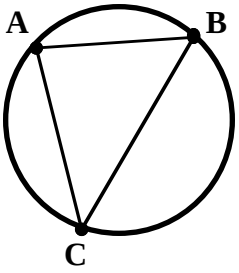


(2) Transposition des axes

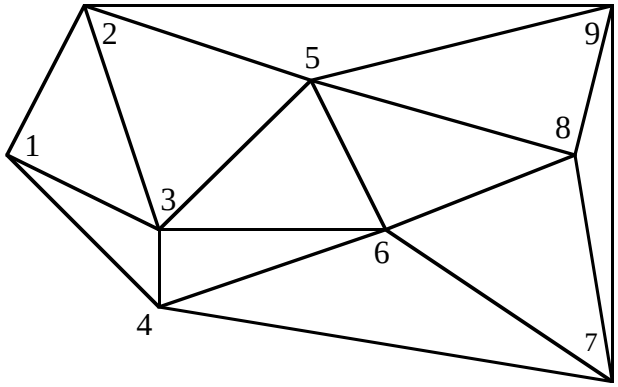
Après avoir lu le fichier des coordonnées des points, le programme présente à l'écran un aperçu (*i.e.*, une carte) de la position des points. L'usager peut alors transposer l'abscisse (ordonnant les points de droite à gauche plutôt que de gauche à droite) ainsi que l'ordonnée (ordonnant les points du haut vers le bas plutôt que du bas vers le haut). Les figures suivantes conserveront l'ordre ainsi établi.

(3) Triangulation de Delaunay

Le critère de la triangulation de Delaunay (Dirichlet, 1850; Upton & Fingleton, 1985) est le suivant. Etant donné trois points A, B et C, le triangle reliant ces trois points sera inclus dans la triangulation si et seulement si le cercle (illustré à gauche) qui passe par ces trois points n'inclut aucun autre point de l'ensemble à l'étude. Ainsi, le fichier de coordonnées (au centre) donnera naissance à la triangulation présentée à droite (N.B. — ne PAS inclure les numéros de points dans VOTRE fichier):



No des points	Coordonnées	
1	0	3
2	1	5
3	2	2
4	2	1
5	4	4
6	5	2
7	8	0
8	7.5	3
9	8	5



Cette triangulation compte les 19 liens suivants:

```
1  4     1  2     1  3     2  3     2  9     3  4     4  7     5  3
5  6     3  6     2  5     5  9     4  6     5  8     6  8     6  7
7  9     7  8     8  9
```

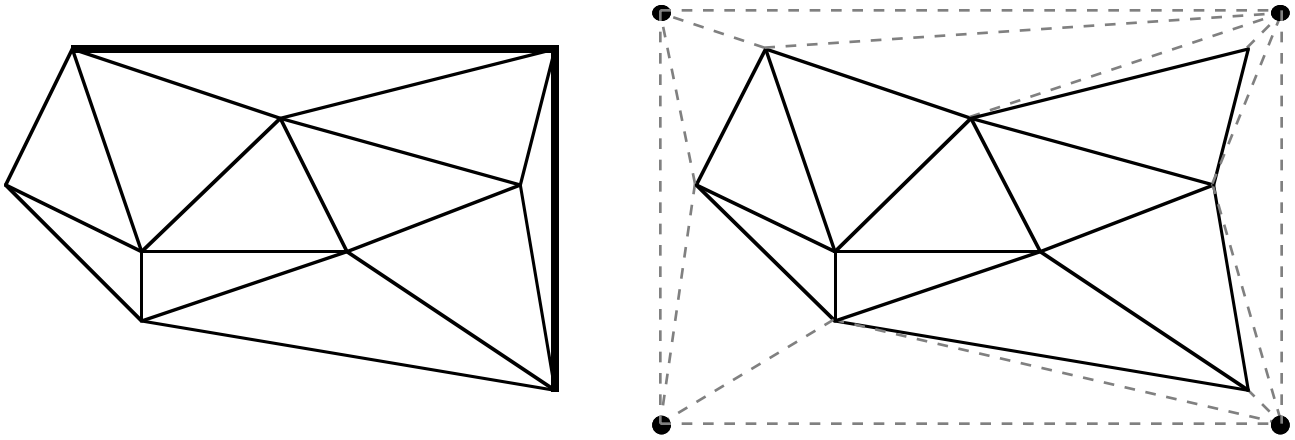
De longs liens peuvent se former en périphérie d'un nuage de points, simplement parce que l'échantillonnage n'inclut pas d'autres points situés plus loin (effet de bordure); par exemple, les liens 2 - 9 et 7 - 9 ci-dessus pourraient ne pas avoir été formés si le nuage de points avait été plus grand. On peut toujours éditer le fichier des liens et éliminer les liens (paires de chiffres) entre objets périphériques trop éloignés. Une autre possibilité consiste à demander au programme de réaliser cette opération pour nous. Pour ce faire, on imposera des "contraintes" au nuage de points. Ces contraintes sont des points supplémentaires, placés à des endroits judicieux et inclus dans l'analyse, dont la présence empêchera la formation des longs liens périphériques indésirables; les liens entre ces points supplémentaires et les vrais points ne seront cependant pas transcrits dans le fichier des liens.

Ces points supplémentaires peuvent être fournis de deux façons différentes. On indiquera la manière choisie en réponse au dialogue proposé par le programme:

```
Contraintes
- Contraintes rectangulaires
- Pas de contrainte
- Contraintes dans le fichier d'entrée
```

Contraintes rectangulaires — Le programme CONNEXIONS contient un algorithme permettant de générer automatiquement des "contraintes" rectangulaires. Pour les points de l'exemple ci-dessus, dont le résultat est présenté de nouveau à gauche ci-dessous, les deux traits gras sont ceux qui seront éliminés par les contraintes. L'algorithme inclut d'abord quatre objets supplémentaires aux coins d'un cadre rectangulaire imaginaire légèrement plus grand que le nuage de points à l'étude; ces objets supplémentaires sont représentés à droite par des points foncés. À la suite du calcul de la triangulation, ces points supplémentaires forment des liens avec les vrais points-objets, et c'est la présence de ces liens (en tirets) qui empêche la formation des deux liens en gras à gauche. Les liens

les points supplémentaires ne sont pas inclus dans la liste des liens.



Contraintes dans le fichier d'entrée — L'usager peut également fournir lui-même comme "contraintes", dans le fichier d'entrée, des points supplémentaires judicieusement disposés; ces points sont décrits dans le fichier par leurs coordonnées en X et en Y, comme les vrais points-objets de l'analyse. Si, par exemple, on avait inclus 6 points supplémentaires de "contrainte" dans le fichier d'entrée *à la suite* des 9 points-objets réels, on aurait dû indiquer au programme qu'il y a 9 vrais points dans l'analyse; puis, en réponse à une question supplémentaire présentée après qu'on ait indiqué que les contraintes se trouvent dans le fichier d'entrée, il aurait fallu dire qu'il y a également 6 points de contrainte dans le fichier.

Pas de contrainte — Aucun point supplémentaire n'est inclus dans le calcul de la triangulation. On pourra toujours éditer le fichier des liens à l'aide d'un éditeur ASCII ou d'un traitement de texte et éliminer à la main les liens (paires de chiffres) entre objets périphériques trop éloignés, s'il y a lieu.

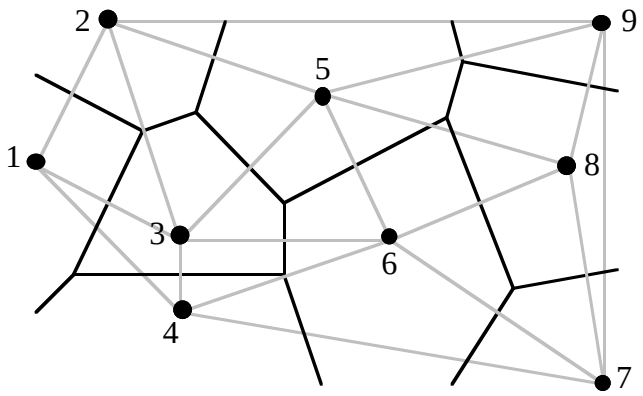
(4) Polygones d'influence

Il peut être intéressant de déterminer la zone d'influence géométrique de chacun des points. La zone d'influence d'un point-objet **A** inclut tous les autres points du plan qui sont plus près de **A** que de tout autre point-objet de l'étude. Les zones d'influence ainsi définies ont la forme de polygones, appelés aussi des *tesselles* (*n.f.*). La figure qu'ils forment s'appelle une *mosaïque* ou un *pavement* (*tessellation* en anglais; adjectif en français: *tessellé*); on s'y réfère souvent comme la *mosaïque de Dirichlet* (1850), les *polygones de Voronoï* (1909) ou les *polygones de Thiessen* (1911), du nom des auteurs qui ont d'abord décrit ces structures mathématiques.

On peut aisément construire ces polygones à partir d'une triangulation de Delaunay, dont ils forment le complément logique. Il suffit en effet de trouver la médiatrice de chaque trait de la triangulation; l'intersection des médiatrices délimite les polygones recherchés. Upton & Fingleton (1985) ainsi que Isaaks & Srivastava (1989) présentent différentes applications de ces mosaïques en analyse spatiale. Le programme offre à l'usager les choix suivants:

```
Choix des traits
- Triangulation seulement
- Polygones seulement
- Triangulation et polygones
```

Pour la figure qui suit, on a choisi l'option "Triangulation et polygones"; la triangulation de Delaunay est en gris et la mosaïque de Dirichlet en noir. Les points-objets, numérotés, se trouvent près du centre des tesselles, mais pas nécessairement en leur centre de masse. La raison en est que la position de la division entre deux tesselles dépend de l'éloignement des plus proches voisins dans cette direction.



L'usager a accès aux différentes options qui se trouvent dans le menu déroulant intitulé "Dessin":

```
Dessin
- Afficher le nombre de liens
- Imprimer le dessin
- Dessiner sur fichier PICT
- Ecrire surfaces
- Terminer
```

Ce menu est accessible après toutes les options de schémas de connexion, mais l'option "Ecrire surfaces" n'est disponible qu'après avoir produit la mosaïque de Dirichlet. C'est pourquoi ce menu est présenté ici.

Afficher le nombre de liens — Le nombre de liens qui ont été écrits dans le fichier est affiché à l'écran. Comme certains programmes qui utilisent ce fichier demanderont à connaître ce nombre, il est prudent de l'inclure dans le nom du fichier de liens.

Imprimer le dessin — Le dessin est imprimé par l'imprimante branchée à l'ordinateur. En particulier, puisque les imprimantes laser et les photocopieuses acceptent les transparents, le dessin pourra être reproduit sur transparent si on désire le superposer à un fond de carte existant.

Dessiner sur fichier PICT — Le contenu du dessin est conservé en format PICT sur un fichier dont le nom est fourni par l'usager. Ce fichier peut être relu par tout programme graphique Macintosh, tel MacDraw, SuperPaint, etc. Le dessin peut donc être édité avant d'être imprimé, ou encore incorporé à un texte (MacWrite, Word, LaserWriter, etc.). Les figures présentées dans cette section ont été en grande partie produites par ce moyen.

Écrire surfaces — Il peut être utile de connaître l'aire de chaque polygone. Ces mesures de surface peuvent être inscrites dans un fichier (ci-dessous), dont le nom est fourni par l'usager. Les surfaces sont dans les mêmes unités (au carré) que les mesures de coordonnées d'origine. Dans certains cas, les polygones périphériques sont fermés, même si leur limite se situe en dehors de la surface reproduite sur le dessin (taille de l'écran). Dans d'autres cas, les polygones périphériques sont ouverts et notés comme tels dans le fichier ci-dessous.

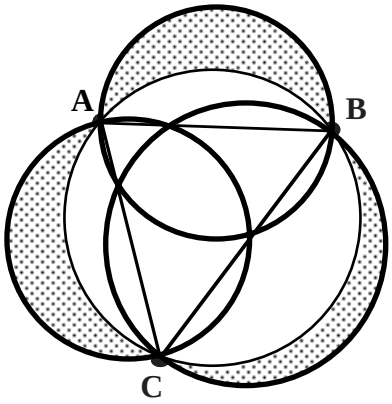
```
1 Ouvert
2 Ouvert
3   4.62500
4 Ouvert
5  15.49844
6  10.06532
7 Ouvert
```

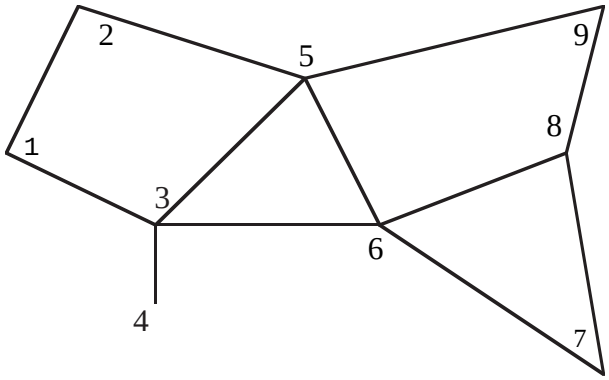
```
8  11.95391
9 Ouvert
```

(5) Schéma de connexion de Gabriel

Le critère du schéma de connexion de Gabriel (Gabriel & Sokal, 1969) diffère de celui de Delaunay, de la façon suivante. Relions deux points **A** et **B** par un trait. Ce trait fera partie du schéma de connexion si aucun autre point **C** ne se trouve à l'intérieur du cercle dont ce trait est le diamètre. En d'autres termes, le trait entre **A** et **B** sera retenu pour faire partie du schéma de connexion si $D^2_{\mathbf{A,B}} < D^2_{\mathbf{A,C}} + D^2_{\mathbf{B,C}}$ pour tout autre point **C** de l'étude ($D^2_{\mathbf{A,B}}$ représente le carré de la distance géographique entre les points **A** et **B**). Une autre façon d'exprimer ce critère est la suivante: si **Centre** représente le point central entre **A** et **B**, alors le trait entre **A** et **B** sera retenu pour faire partie du schéma de connexion si $D_{\mathbf{A,B}}/2 < D_{\mathbf{Centre,C}}$ pour tout autre point **C** de l'étude.

L'exemple ci-dessous (à droite) représente le schéma de connexion de Gabriel pour les mêmes points que dans l'exemple de la triangulation de Delaunay. On peut voir que les 12 traits formant le schéma de Gabriel sont un sous-ensemble des 19 traits retenus pour la triangulation de Delaunay (ci-dessus). En effet, comme on le voit dans le schéma ci-dessous (gauche), les cercles (en gras) correspondant au critère de Gabriel peuvent inclure, dans les zones ombrées du côté extérieur au cercle de Delaunay (ligne fine), certains points-objets que le cercle du critère de Delaunay n'inclut pas, si bien que certains traits autorisés par le critère de Delaunay seront exclus par celui de Gabriel.

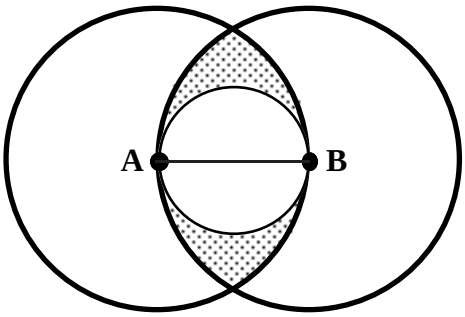


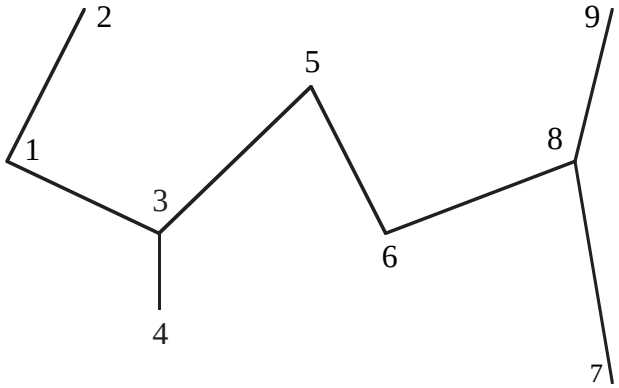


Dans cette option, il n'est pas utile de demander si on désire imposer une "contrainte" (voir ci-dessus), car les longs traits qui pourraient se former en périphérie de l'ensemble de points sont automatiquement éliminés par le critère de Gabriel.

(6) Schéma de voisinage relatif

Le critère de voisinage relatif (*"relative neighborhood graph"*, en anglais) diffère de celui de Gabriel de la façon suivante. Relions deux points **A** et **B** par un trait; traçons un cercle centré en **A** et un second centré en **B**, ces deux cercles ayant pour rayon le trait de **A** à **B**. Ce trait fera partie du schéma de connexion si aucun autre point **C** de l'étude ne se trouve inclus dans l'intersection de ces deux cercles. En d'autres termes, le trait entre **A** et **B** sera retenu pour faire partie du schéma de connexion si $D_{\mathbf{A,B}} \leq \max(D_{\mathbf{A,C}}, D_{\mathbf{B,C}})$ pour tout autre point **C** de l'étude. L'exemple ci-dessous (à droite) représente le schéma de voisinage relatif pour les mêmes points que dans les exemples ci-dessus. On peut voir que les 8 traits (8 = nombre d'objets - 1) formant le schéma de voisinage relatif sont un sous-ensemble des 12 traits formant le schéma de Gabriel (ci-dessus). En effet, comme on le voit dans le schéma ci-dessous (gauche), l'intersection des deux cercles (en gras) formant le critère de voisinage relatif peut inclure, dans la partie ombrée, des points-objets que le cercle de Gabriel (petit cercle, ligne fine) n'inclut pas, si bien que certains traits autorisés par le critère de Gabriel seront exclus par celui du voisinage relatif.


A
B


1
2
3
4
5
6
7
8
9

*CONVERSION*Macintosh *ou CONVERT*CMS/VMS

Que fait CONVERSION ?

Ce programme utilitaire permet de convertir une matrice de similarité (S) de type SIMIL en une matrice de distance (D) ou vice-versa. Les versions CMS et VMS n'utilisent que la formule

$S_{ij} = 1 - D_{ij}$ ou $D_{ij} = 1 - S_{ij}$

alors que la version Macintosh permet également de convertir à l'aide des formules

$S_{ij} = \sqrt{[1 - D_{ij}]}$ ou $D_{ij} = \sqrt{[1 - S_{ij}]}$

et $S_{ij} = 1 - [(D_{ij} - Dmin)/(Dmax - Dmin)]$ ou $D_{ij} = 1 - [(S_{ij} - Smin)/(Smax - Smin)]$

Ce programme a été écrit parce que la plupart des programmes de "R" requièrent, pour le fichier SIMIL qui leur est fourni, qu'il soit du type SIMILARITÉS. Les fichiers de distance doivent donc dans bien des cas être convertis; dans le cas où les distances sont plus grandes que 1, la première forme de la transformation produira des "similarités" négatives, mais les programmes subséquents sont conçus de façon à traiter celles-ci convenablement. La dernière forme de la transformation garantit que les similarités obtenues seront obligatoirement bornées entre 0 et 1.

Fichiers d'entrée et de sortie

(1) Fichier d'entrée

Le fichier d'entrée du programme CONVERSION est un fichier binaire de type SIMIL. Il contient une matrice de similarité ou de distance produite soit par SIMIL, ou encore par IMPORT-EXPORT (version Macintosh) ou IMPORT (versions CMS et VMS).

(2) Fichier de sortie

La sortie de CONVERSION est aussi un fichier binaire de type SIMIL contenant la matrice transformée. Le fichier binaire converti porte la mention "(CONVERT)" dans le bloc d'informations qui est imprimé automatiquement par plusieurs des programmes. On peut avoir recours par exemple au programme REGARDE, qui permet de lire ces informations ainsi que le contenu des fichiers de type SIMIL:

```
TITRE:  Matrice de similarites
DATE 20/02/91
FONCTION  s15
(CONVERT)
NOMBRE D' OBJETS : 57
NOMBRE DE DESCRIPTEURS :  3
```

Exemple

(Les réponses de l'utilisateur sont en **caractères gras**)

```
Quel est le nom du fichier SIMIL a transformer S-D ou D-S ?
(Par defaut: "... data a")
fichier  s15

Quel nom doit recevoir le fichier produit par ce programme?
(Par defaut: "... data a")
fichier  dist

Execution begins...

 P R O G R A M M E   C O N V E R T
 Convertit une matrice S en D ou une matrice D en S
 VERSION 3.0b
 TRANSFORMATION SIMILARITE <-> DISTANCE
 AUTEUR: A. VAUDOR

 Fin du programme.
```

*DISTANCES GÉOGRAPHIQUES*Macintosh *ou DIST*CMS/VMS

Que fait DISTANCES GÉOGRAPHIQUES ?

À partir d'un fichier de coordonnées cartographiques de localités, ce programme calcule les distances géographiques entre ces localités *en suivant la courbure de la terre*.

Fichiers d'entrée et de sortie

(1) Fichier des coordonnées (entrée)

Le fichier des coordonnées est un fichier ASCII rectangulaire, où les lignes correspondent aux stations alors que les colonnes correspondent aux coordonnées *latitude* et *longitude* (avec la *latitude* en premier). Il peut être présenté de l'une des façons décrites à la section des **Options** ci-dessous.

Notez par ailleurs que si les points chevauchent le 0 degré de longitude, les longitudes à l'ouest de Greenwich peuvent être présentées soit avec un signe négatif (-), soit sur 360 degrés. La latitude des points au sud de l'Équateur doit être présentée avec un signe négatif (-). Évidemment, si tous les points sont à l'ouest de Greenwich, ou encore dans l'hémisphère sud, le signe devient inutile.

(2) Matrice de distances (sortie)

Le fichier contenant la matrice de distances est en ASCII, c'est-à-dire en caractères lisibles. La matrice de distances est carrée, avec des zéros sur la diagonale. De plus, dans la version Macintosh, la première ligne du fichier de sortie reproduit le nom du fichier d'entrée (fichier des coordonnées). Pour en faire une matrice binaire de type SIMIL, il faudra transformer cette matrice à l'aide de IMPORT (versions CMS et VMS) ou de IMPORT-EXPORT (version Macintosh). Les distances peuvent être exprimées de l'une ou l'autre des façons décrites à la section des **Options** ci-dessous.

Options

Les options disponibles pour le fichier d'entrée sont décrites par le menu suivant, présenté par le programme aux usagers:

```
0: degrés décimaux
1: degrés point minutes (ex.: 35.04)
2: degrés espace minutes (ex.: 35 04)
3: degrés point minutes point secondes (ex.: 35.04.05)
4: degrés espace minutes espace secondes (ex.: 35 04 05)
```

Quant au fichier de sortie, les options suivantes sont disponibles:

```
0: distances en radians
1: distances en degrés
2: distances en milles marins (ou minutes d'arc)
3: distances en milles
4: distances en kilomètres
```

EXPNTS[CMS] **Que fait EXPNTS ?** À partir d'un fichier contenant une matrice de distances produite par SIMIL, ce programme permet de créer un nouveau fichier binaire contenant une matrice de distances, utilisable par le logiciel d'analyses multidimensionnelles NT-SYS (Numerical Taxonomy SYSteme : Rohlf *et al.*, 1971). Le progiciel NT-SYS contient dans son programme de cadrage multidimensionnelle non-métrique (programme MDSCALE). Puisque cette analyse très utile n'est pas disponible dans le progiciel "R", le programme EXPNTS permet de transférer au NT-SYS, sur grands ordinateurs IBM, les matrices de ressemblance (similarités ou distances) calculées à l'aide du progiciel "R". En effet, plusieurs mesures de ressemblance disponibles dans le programme SIMIL de "R" ne sont pas disponibles dans NT-SYS. Pour l'utilisation de NT-SYS en version MS-DOS, on aura plutôt recours à l'utilitaire EXPORT (versions CMS et VMS de "R") ou IMPORT-EXPORT (version Macintosh), puisque la version MS-DOS de NT-SYS n'utilise que des matrices écrites en ASCII.

Le progiciel NT-SYS, développé par le Prof. F. James Rohlf, est distribué par *Exter Software Inc.*, 100 North Country Road, Bldg. B, Setauket, New York 11733, USA (versions disponibles: pour machines MS-DOS et pour grands ordinateurs).

Fichiers d'entrée et de sortie

(1) Fichier d'entrée

Le fichier d'entrée de EXPNTS est un fichier binaire structuré produit par SIMIL, IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh); la structure des matrices de ressemblance binaires est décrite au chapitre du programme SIMIL. C'est donc le fichier de sortie du programme SIMIL qui sert de fichier d'entrée à EXPNTS. Ce fichier contient la matrice de ressemblance (similarités ou distances) écrite en binaire. Le programme EXPNTS lira lui-même sur le fichier d'entrée l'information quant au nombre d'objets que contient cette matrice.

(2) Fichier de sortie

Le fichier de sortie est un nouveau fichier binaire contenant une matrice de distances, écrite sous une forme compatible avec le progiciel NT-SYS. Il contient la même matrice triangulaire supérieure de distances que le fichier d'entrée, mais des termes diagonaux égaux à zéro lui sont ajoutés. Ce fichier ne contient aucune en-tête, contrairement aux fichiers de type SIMIL. Il n'est pas possible de relire cette matrice binaire grâce à REGARDE, ni à l'aide d'un éditeur ASCII ou d'un traitement de texte.

Questions du programme

Le programme d'appel demande simplement quel est le nom du fichier d'entrée et du fichier de sortie. Le programme lui-même ne pose qu'une seule question: "`Transformation Similarités-Distances ?`" — Si le fichier d'entrée contient une matrice de similarités, cette question offre la possibilité de transformer les similarités en distances ($D = 1 - S$), puisque le programme NT-SYS s'attend à ce que son fichier d'entrée contienne une matrice de distances.

*EXPORT*CMS/VMS

Que fait EXPORT ?

Ce programme permet de transformer des matrices de ressemblance binaires produites par le programme SIMIL en des matrices carrées écrites en ASCII (caractères lisibles). EXPORT remplit l'une des fonctions du programme IMPORT-EXPORT de la version Macintosh. De telles matrices carrées écrites en caractères lisibles peuvent être utiles,

- soit pour les présenter dans des publications,
- soit pour les passer à des programmes appartenant à d'autres progiciels,
- soit encore s'il est nécessaire de transférer des matrices de ressemblance d'un type d'ordinateur à un autre, les matrices binaires produites par SIMIL n'étant pas transférables entre machines ayant des représentations différentes pour les nombres en points flottants. Les matrices carrées produites par ce programme peuvent être relues par le programme IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh) s'il est nécessaire de les reconvertir en format binaire de type SIMIL.

Fichiers d'entrée et de sortie

Le fichier d'entrée est une matrice de ressemblance (similarités, distances ou coefficients de dépendance entre variables) en format SIMIL. La structure des matrices de ressemblance binaires est décrite au chapitre du programme SIMIL. Le fichier de sortie est une matrice carrée en caractères lisibles (ASCII) écrit en format (8F10.7), avec des 1.0000000 sur la diagonale. Ces valeurs peuvent aisément être converties en des 0.0000000 à l'aide d'un éditeur ASCII, si cela est nécessaire.

Questions du programme

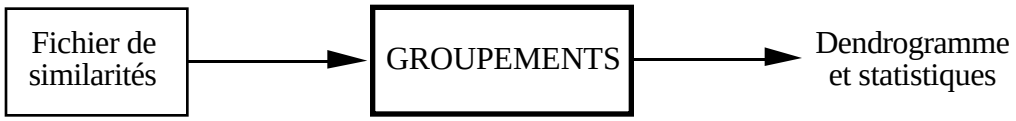
Le programme d'appel demande simplement quel est le nom du fichier d'entrée et du fichier de sortie. Le programme lui-même ne pose aucune question à l'usager. Voir la fin de la section portant sur le programme IMPORT-EXPORT, où est présenté un exemple de matrice carrée ASCII produite en sortie.

GROUPEMENTS Macintosh

Que fait GROUPEMENTS ?

Le programme GROUPEMENTS réalise des groupements agglomératifs selon une gamme de méthodes, décrites succinctement ci-dessous. Ce programme remplit pour la version Macintosh le même rôle que les programmes INTERLNK et LANCE des versions CMS et VMS.

Fichiers d'entrée et de sortie



Le fichier d'entrée est une matrice de similarités de type SIMIL. La sortie est un dendrogramme, accompagné ou non de certaines statistiques (voir ci-dessous); il peut être envoyé soit à une imprimante, soit à un fichier. Si on dispose d'une imprimante laser, les dendrogrammes produits sont de haute qualité graphique et peuvent être inclus directement dans des publications; l'usager peut décider de leur taille (largeur en cm) ainsi que des polices de caractères employées pour les réaliser. Les dendrogrammes inscrits dans un fichier ont la même facture que ceux produits par INTERLNK et LANCE des versions CMS et VMS (leur largeur est comptée en nombre de caractères d'imprimerie).

Options

On demande d'abord à l'usager de choisir entre l'algorithme de Lance & Williams (1966a, 1967), utilisé également par le programme LANCE des versions CMS et VMS, et l'algorithme à liaison proportionnelle (Sneath, 1966), employé aussi par le programme INTERLNK.

```
Type de calcul:
  Lance & Williams
  Liens intermédiaires
```

Si on a choisi l'algorithme à liaison proportionnelle ("Liens intermédiaires"), on doit fournir la connexité désirée ("Proportion des liens"). Par contre, si on préfère l'algorithme général de groupement agglomératif de Lance & Williams, le choix suivant est offert:

```
Type de groupement:
  Association moyenne (UPGMA)
  Poids proportionnels (WPGMA)
  Groupement centroïde (UPGMC)
  Groupement médian (WPGMC)
  Méthode de Ward
  Autre
```

Si on choisit l'option "Autre", on doit encore fournir la valeur des paramètres *alpha*[j], *alpha*[m], *bêta* et *gamma* requis par cet algorithme; voir la description du programme LANCE. Enfin, après chaque groupement, l'usager doit encore répondre à deux questions pour indiquer s'il désire, ou non, obtenir les statistiques complémentaires au groupement (voir leur description ci-dessous):

```
Chaîne des liens primaires?                                  [oui ou non]
Corrélations cophénétiques, distance de Gower et entropie?   [oui ou non]
```

On peut enfin faire écrire la matrice des distances cophénétique dans un fichier de type SIMIL.

Statistiques de groupement

Les statistiques suivantes sont disponibles pour permettre de juger l'adéquation qui existe entre le groupement produit et la matrice de similarité d'origine.

(1) La chaîne des liens primaires

La *chaîne des liens primaires,* ou *dendrites*, porte aussi le nom de *réseau, réseau de Prim*, *squelette arborescent* ou *arbre de longueur minimum*, (*minimum spanning tree, minimum length tree* ou *shortest spanning tree* en anglais) est l'ensemble des liens entre paires d'objets qui représente la structure fondamentale du groupement. Un *lien primaire* est défini formellement comme le premier lien qui rend un objet membre d'un groupe, ou encore qui produit la fusion de deux groupes, dans le cas du groupement à liens simples (Legendre & Legendre, 1984a). Dans les programmes de groupement agglomératif du progiciel "R", la notion de lien primaire est ici étendue pour représenter, lors de la fusion de deux groupes, le lien de similarité unissant les deux objets (un dans chacun des deux groupes) qui sont les plus près l'un de l'autre (*i.e.*, plus forte similarité); la chaîne des liens primaires sera donc l'ensemble de ces premiers liens de similarité, quelle que soit la méthode ayant donné naissance aux groupes.

Reprenant l'exemple des 5 mares utilisé par Legendre & Legendre (1984a) comme exemple dans leur chapitre sur le groupement, un groupement à liens simples a été réalisé. Le dendrogramme est présenté dans la section portant sur le programme INTERLNK. La chaîne des liens primaires fournie par le programme pour le groupement à liens simples est la suivante:

```
L i e n s   p r i m a i r e s

    Niveau   Distance       Chaîne
-----------------------------------------------

   0.40000   0.40000  (MARE 214   , MARE 212   )
   0.78600   0.78600  (MARE 432   , MARE 214   )
   0.70000   0.70000  (MARE 431   , MARE 233   )
   0.50000   0.50000  (MARE 432   , MARE 431   )
```

Cette liste signifie qu'au niveau de groupement D = 0.4, la chaîne des liens primaires s'enrichit d'un premier lien entre les mares qui portent le nom de 212 et 214; ces deux objets sont situés à la distance D = 0.4 (ou S = 1 - 0.4 = 0.6) dans la matrice de similarités d'origine; et ainsi de suite. Dans le cas du groupement à liens simples, les niveaux de fusion sont toujours égaux aux distances entre les objets les plus proches, par définition de la méthode. Dans d'autres types de groupement, tel n'est pas le cas. Ainsi, dans le groupement selon l'association moyenne (UPGMA: ci-dessous), les deux valeurs diffèrent sur la seconde et la troisième lignes; le 'niveau' est toujours le niveau de fusion des groupes dans le dendrogramme, alors que la 'distance' est le complément de la valeur de similarité (D = 1 - S) entre les deux objets les plus voisins des deux groupes:

```
L i e n s   p r i m a i r e s

    Niveau   Distance       Chaîne
-----------------------------------------------

   0.40000   0.40000  (MARE 214   , MARE 212   )
   0.94200   0.78600  (MARE 432   , MARE 214   )
   0.75000   0.70000  (MARE 431   , MARE 233   )
   0.50000   0.50000  (MARE 432   , MARE 431   )
```

Notez que dans la chaîne des liens primaires, tout comme dans le dendrogramme, les niveaux de fusion et les ressemblances entre objets formant la chaîne sont exprimés en distances et non en

fusion et les ressemblances entre objets formant la chaîne sont exprimés en distances et non en similarités; ce choix résulte du fait que pour des coefficients de distance prenant des valeurs dans l'intervalle [0, ∞], les dendrogrammes pourraient avoir des niveaux de fusion négatifs s'ils étaient exprimés en similarités S = (1 - D); cela n'aurait pas d'effet sur le groupement, mais rendrait la lecture de ces niveaux de fusion plus difficile pour l'utilisateur.

(2) Les corrélations cophénétiques

Tout groupement hiérarchique peut être représenté par une *matrice cophénétique* entre objets (Sokal & Rohlf, 1962; Legendre & Legendre, 1984a; Jain & Dubes, 1988; etc.). Dans cette matrice, la similarité entre deux objets est égale à la valeur du niveau de fusion qui permet de joindre les deux objets en question dans le dendrogramme. Pour toutes les méthodes de groupement agglomératif — sauf (occasionnellement) pour le groupement centroïde, le groupement médian, ainsi que les groupements obtenus par l'emploi de certaines combinaisons non habituelles des paramètres du modèle général de groupement agglomératif de Lance & Williams (1966, 1967) — la matrice cophénétique ainsi obtenue représente une ultramétrique.

La corrélation linéaire entre la matrice cophénétique et la matrice de similarités d'origine (en excluant la diagonale) porte le nom de *corrélation cophénétique*, *corrélation de matrices* ou *statistique de Mantel centrée-réduite*. Cette corrélation mesure à quel point le groupement correspond à la matrice de similarités d'origine, puisqu'un groupement qui rendrait totalement compte des similarités de la matrice d'origine produirait une corrélation cophénétique de 1. Notez que cette corrélation ne peut, logiquement, être testée quant à sa signification statistique, puisque la matrice cophénétique n'est pas indépendante de la matrice de similarités d'origine, étant issue d'elle *via* l'algorithme de groupement. Pour tester cette corrélation, il faudrait prétendre que les deux matrices sont indépendantes l'une de l'autre sous H_0; autrement dit, il faudrait admettre comme probable que l'algorithme de groupement ait une efficacité nulle, ce qui n'est pas le cas des algorithmes de groupement courants…

Si on trace un graphique (diagramme de Shepard) des similarités (ou des distances) de la matrice cophénétique en fonction des similarités (ou distances) d'origine, il peut arriver que la relation soit curviligne plutôt que linéaire. Si on s'intéresse davantage à la structure topologique du dendrogramme qu'à la longueur exacte de ses branches, il convient de rechercher une relation monotone plutôt qu'une relation nécessairement linéaire entre les deux matrices. Dans ce cas, le calcul d'une corrélation non-paramétrique sera approprié, plutôt qu'une corrélation linéaire de Pearson; la corrélation non-paramétrique de Kendall (τ_b) entre la matrice cophénétique et la matrice d'origine est fournie par le programme. Les coefficients de corrélation prennent des valeurs qui se situent dans l'intervalle [-1, 1]. On s'attend à ce que le signe de la corrélation cophénétique soit positif, puisque la comparaison se fait entre la matrice des similarités d'origine et la matrice de similarités cophénétiques. Plus la valeur de la corrélation cophénétique est élevée, plus l'ajustement est bon. Voici un exemple des mesures d'ajustement fournies par le programme, pour le groupement par la méthode de l'association moyenne (UPGMA) des 5 mares:

```
Corrélations cophénétiques
      tau b de Kendall    0.77364
      r de Pearson        0.95111
      distance de Gower   0.03962
```

La distance de Gower, la dernière mesure d'ajustement disponible dans les programmes de groupement agglomératif de "R", est décrite ci-dessous.

(3) La distance de Gower

La distance de Gower (1983) est la somme des carrés des écarts entre les valeurs de la matrice de similarités cophénétiques et des similarités d'origine. Cette mesure d'ajustement prend des valeurs qui

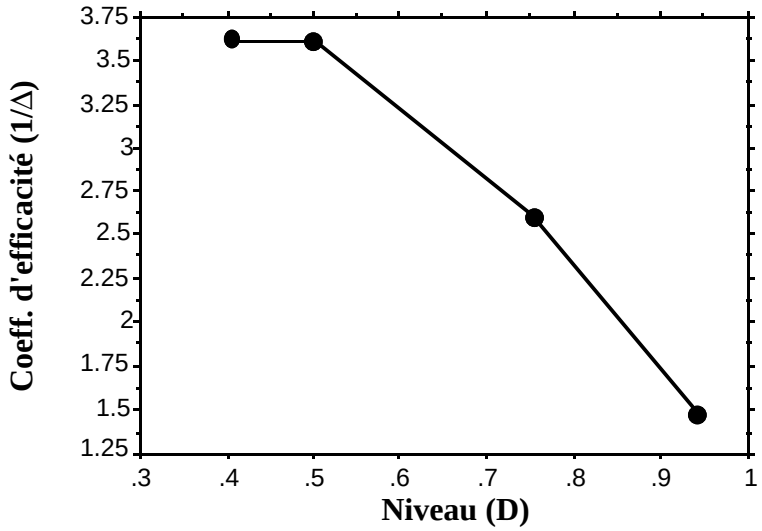
se situent dans l'intervalle [0, ∞]. Plus la distance de Gower est faible, plus l'ajustement est bon. Comme c'est le cas pour les corrélations cophénétiques, cette mesure a une valeur comparative entre des groupements obtenus à partir d'une même matrice de similarités.

(4) Les coefficients d'efficacité

Les *coefficients d'efficacité* (Lance & Williams, 1966b) se calculent par 1/Δ, où Δ (delta) représente la valeur de réduction de l'information dans la classification, réduction produite par la fusion de groupes. Cette réduction se calcule comme l'entropie de la classification avant la fusion, moins l'entropie après la fusion. Un coefficient d'efficacité est fourni par le programme pour chaque niveau de fusion. Lorsque l'algorithme ne regroupe que des objets individuels ou encore de petits groupes, les valeurs de Δ sont faibles; par conséquent, les valeurs correspondantes du coefficient d'efficacité sont fortes. Si on trace un graphique des valeurs du coefficient d'efficacité en fonction des étapes du groupement agglomératif, les minima de ce graphique indiquent quelles sont les fusions les plus importantes. Si on cherche à sélectionner un point de coupure dans le dendrogramme, le coefficient d'efficacité peut aider à prendre cette décision. Il ne s'agira cependant en aucun cas d'un critère de décision "obligatoire" puisqu'aucun test de signification statistique n'a été réalisé.

```
Coefficients d'efficacité

           Niveau  Entropie    delta    1/delta
-------------------------------------------------
          0.00000   1.60944
          0.40000             0.27726   3.60674
                    1.33218
          0.50000             0.27726   3.60674
                    1.05492
          0.75000             0.38191   2.61843
                    0.67301
          0.94200             0.67301   1.48586
                    0.00000
```

Le graphique suivant présente les valeurs du coefficient d'efficacité (1/Δ) en fonction des niveaux de fusion (distances). Dans cet exemple simple, la partition la plus "efficace" est la dernière. La meilleure coupure verticale dans le dendrogramme se situerait donc avant ce dernier niveau, ce qui produirait deux groupes.



IMPORTCMS/VMS

Que fait IMPORT ?

Ce programme permet d'importer des matrices de ressemblance, les transformant du format ASCII au format binaire SIMIL requis pour qu'elles puissent être lues par les autres programmes de "R". IMPORT remplit l'une des fonctions du programme IMPORT-EXPORT de la version Macintosh. Les matrices carrées en caractères lisibles peuvent avoir été écrites par des programmes appartenant à d'autres progiciels. Il peut également s'agir de matrices de ressemblance calculées par SIMIL sur un autre type d'ordinateur, matrices qui auront été converties au format ASCII par le programme EXPORT avant d'être transférées d'une machine à l'autre, pour être reconverties au format binaire de type SIMIL à l'aide de IMPORT. Enfin, dans certains cas (études du comportement, sociologiques, de génétique moléculaire, etc.), les données brutes se présentent sous la forme de matrices d'association entre individus; de telles matrices peuvent être saisies sur fichier à l'aide d'un éditeur ASCII, puis importées grâce au programme IMPORT pour être analysées par les programmes de "R".

Fichiers d'entrée et de sortie

Le fichier d'entrée contient une matrice carrée écrite en ASCII (caractères lisibles), produite de l'une des façons décrite ci-dessus. Le fichier de sortie est un fichier binaire de type SIMIL contenant les mêmes informations. La structure des matrices de ressemblance binaires est décrite au chapitre du programme SIMIL.

Questions du programme

Après que le programme d'appel ait demandé le nom des fichiers d'entrée et de sortie, le programme lui-même demande:

```
TAILLE DE LA MATRICE (nombre d'objets ou de variables)
```

On répond à cette question par un seul nombre entier, correspondant au nombre d'objets (en mode Q) ou de descripteurs (en mode R) qui sont comparés dans ladite matrice. La question suivante est:

```
NOMBRE INITIAL D'OBJETS s'il s'agit d'une matrice de corrélations;
Dans le cas contraire, donnez de nouveau la taille de la matrice.
```

En dernier lieu, le programme demande un titre. Ces informations servent d'abord à compléter le bloc d'informations joint d'office à tout fichier de type SIMIL (voir l'exemple dans la description du programme IMPORT-EXPORT, ci-dessous). L'information concernant le nombre d'objets, pour une matrice de ressemblance calculée en mode R (matrice de covariance ou de corrélation), est nécessaire pour certains autres programmes, tel notre programme de calcul de corrélations partielles (non inclus pour le moment dans la version de "R" qui vous est fournie), afin que les tests de signification soient réalisés de façon adéquate. Dans les autres cas, le nombre donné en réponse à cette question sera simplement inscrit dans le bloc d'informations, mais il ne sera pas utilisé par les programmes d'analyse subséquents. Notez que dans le bloc d'informations, la FONCTION indiquera que le fichier binaire de type SIMIL a été produit par le programme IMPORT.

IMPORT-EXPORT Macintosh

Que fait IMPORT-EXPORT ?

Ce programme permet d'importer des matrices de ressemblance, les transformant du format ASCII au format binaire SIMIL nécessaire pour qu'elles puissent être lues par les autres programmes de ce progiciel. IMPORT-EXPORT peut également réaliser l'opération contraire et transformer des matrices de ressemblance du format SIMIL au format ASCII. Ce programme Macintosh remplit les mêmes fonctions que les programmes IMPORT et EXPORT des versions CMS et VMS.

Fichiers d'entrée et de sortie

Les matrices de ressemblance en format SIMIL ou en format ASCII peuvent être employées soit comme fichier d'entrée de ce programme, soit comme fichier de sortie. La structure des matrices de ressemblance binaires est décrite au chapitre du programme SIMIL. Les matrices de ressemblance en ASCII peuvent se présenter sous différentes formes, décrites ci-dessous.

Options

La première question posée par le programme concerne le type de conversion demandé, soit de la gauche vers la droite dans le schéma ci-dessus, ou de la droite vers la gauche:

```
Type de conversion
  De fichier caractères à fichier SIMIL
  De fichier SIMIL à fichier caractères
```

Si on a choisi de convertir un fichier binaire de type SIMIL en fichier ASCII (en caractères lisibles), le programme ne demande que le nom du fichier d'entrée ainsi que le nom que l'on désire attribuer au fichier de sortie en caractères. Ce fichier ASCII sera une matrice carrée symétrique avec des 1.0000000 sur la diagonale. Ces valeurs peuvent aisément être converties en des 0.0000000 à l'aide d'un éditeur ASCII, si cela est nécessaire.

Si on a plutôt choisi de convertir un fichier ASCII, on doit décrire d'abord la taille de la matrice (*i.e.*, le nombre d'objets ou de descripteurs qui y sont comparés), puis sa forme:

```
Fichier d'entrée:                    [Dans le menu: donner le nom du fichier d'entrée]
  Taille de la matrice                [On répond par un seul nombre entier]

Type de matrice:
  Carrée avec diagonale
  Carrée sans diagonale
  Triangulaire supérieure avec diagonale
  Triangulaire supérieure sans diagonale
```

Le programme demande enfin de fournir deux informations qui auraient été disponibles si les matrices de ressemblance avaient été calculées par le progiciel SIMIL, soit la taille du tableau de

données d'origine dans l'autre dimension, ainsi que le titre que l'usager veut bien joindre à cette matrice:

```
Nombre d'objets (en mode Q) ou de descripteurs (en mode R)        [Un entier]
Titre de ce travail                  [On répond par un titre d'au plus 80 caractères]
```

Explication — Soit un tableau de données de *n* lignes et *p* colonnes. Si les mesures de ressemblance contenues dans la matrice à transformer ont été calculées entre les lignes de ce tableau, alors le programme veut maintenant connaître le nombre de colonnes. Si au contraire c'est entre les colonnes que cette matrice de ressemblance a été calculée, il faut maintenant fournir le nombre de lignes du tableau de données d'origine. Cette information sert d'abord à compléter le bloc d'informations joint d'office à tout fichier de type SIMIL:

```
FICHIER D'ENTREE: Fichier d'entrée
     TITRE: Hydrologie des lacs de la Baie de James
     DATE: 2/5/91
     FONCTION:  (ImpExp)
     Nombre d'objets: 32
     Nombre de descripteurs: 10
```

(Notez que la FONCTION indique que ce fichier binaire a été produit par le programme IMPORT-EXPORT.) L'information concernant le nombre d'objets, pour une matrice de ressemblance calculée en mode R (matrice de covariance ou de corrélation), est nécessaire pour certains autres programmes, tel notre programme de calcul de corrélations partielles (non inclus pour le moment dans la version de "R" qui vous est fournie), afin que les tests de signification soient réalisés de façon adéquate.

Exemple

La matrice triangulaire supérieure suivante, sans diagonale, mesure les distances routières en km entre 6 villes du Québec. Les chiffres ont été alignés pour en faciliter la lecture, mais cela n'est pas requis par le programme (lecture en format libre). Les valeurs peuvent être des nombres entiers, ou encore des nombres réels avec ou sans chiffre précédant le point décimal (.138, 0.138 ou -.57 sont admis).

```
198 368  57  882 311
    549 238 1063 482
        311  517  80
             824 253
                 594
```

Après transformation en matrice binaire (qui ne peut être illustrée ici), puis retour en matrice ASCII, on obtient le résultat suivant:

```
1.0000000  198.00000 368.00000  57.00000  882.00000 311.00000
198.00000  1.0000000 549.00000 238.00000 1063.00000 482.00000
368.00000  549.00000 1.0000000 311.00000  517.00000  80.00000
 57.00000  238.00000 311.00000 1.0000000  824.00000 253.00000
882.00000 1063.00000 517.00000 824.00000  1.0000000 594.00000
311.00000  482.00000  80.00000 253.00000  594.00000 1.0000000
```

Bien entendu, dans cette matrice de distances routières, les "1" sur la diagonale n'ont pas de sens et devront être remplacés par des zéros, s'ils nuisent à la suite du calcul. La valeur "1" a été choisie parce qu'elle est la valeur appropriée dans deux situations courantes: d'abord pour toutes les matrices de similarités, ensuite pour les matrices de corrélations. Les 1.0000000 sont facilement identifiables pour l'éditeur car ce sont les seules valeurs à 7 décimales dans le fichier de sortie ASCII.

*INTERLNK*CMS/VMS

Que fait INTERLNK ?

Ce programme réalise un groupement agglomératif à liaison proportionnelle (liens intermédiaires). La connexité du groupement (*Co*), qui est fixée par l'usager, peut varier de 0 à 100%, ce qui représente toute la gamme des solutions depuis les liens simples (*Co* = 0) jusqu'aux liens complets (*Co* = 1). À connexité voisine de 50%, le groupement respecte approximativement les propriétés métriques de l'espace de référence; à connexité faible, il est sujet à l'enchaînement (contraction de l'espace de référence), alors qu'à connexité élevée se produit le phénomène inverse, soit la dilatation apparente de l'espace autour des noyaux de groupement (Lance & Williams, 1967).

Le programme d'appel INTERLNK lance à tour de rôle trois programmes différents: (1) un programme de tri qui récrit la matrice de similarités en ordre de similarités décroissantes, (2) le programme de groupement lui-même, et finalement (3) le programme qui trace le dendrogramme. L'usager peut demander à ce dernier programme de calculer différentes statistiques (chaîne des liens primaires, corrélations cophénétiques, distance de Gower, coefficients d'efficacité) qui ont été décrites à la fin de la section portant sur le programme GROUPEMENTS.

Fichiers d'entrée et de sortie

(1) Le fichier d'entrée

Le fichier d'entrée doit impérativement être un fichier de similarités, et NON PAS de distances, écrit par les programmes SIMIL ou IMPORT; INTERLNK n'existe qu'en versions CMS et VMS. Une matrice de distances peut aisément être convertie en une matrice de similarités à l'aide du programme utilitaire CONVERT.

Le nombre maximum d'objets qui peuvent être traités par ce programme, ainsi que le nombre maximum de groupes simultanés, sont fixés par les paramètres MAXDIM et MAXGROUPES respectivement, en début du programme. Ces paramètres peuvent être ajustés à la taille des problèmes à traiter, avant la compilation.

(2) Le fichier de sortie

Ce fichier contient le dendrogramme décrivant le groupement agglomératif, ainsi que les statistiques de groupement. Celles-ci sont décrites en détail à la section relative au programme GROUPEMENTS. Si on a attribué des noms aux objets dans le fichier de données brutes soumis à SIMIL (10 premiers caractères), le dendrogramme présente ces identificateurs, au lieu des numéros d'ordre que le programme attribuera autrement aux objets.

Questions du programme

Après que le programme d'appel ait demandé le nom des fichiers d'entrée et de sortie, le programme de groupement lui-même demande seulement quel degré de connexité (*Co*) devra être employé par l'algorithme de groupement à liaison proportionnelle.

```
PROPORTION DES LIENS (Connexité) ?
```

La connexité varie de 0 pour les liens simples à 1.0 pour les liens complets. On répond à cette question par un nombre réel compris entre 0 et 1. Pour les nombres fractionnaires, le langage PASCAL exige que l'on écrive par exemple "0.75" et non ".75".

Les questions suivantes sont posées par le programme DENDRO qui se charge de tracer le dendrogramme et de calculer les statistiques de groupement; voir la description de ces statistiques à la section portant sur le programme GROUPEMENTS. La largeur du dendrogramme qui sera dessiné est fixée par l'utilisateur, qui doit déterminer le <u>nombre de caractères d'imprimerie</u> qui seront utilisés pour tracer le dendrogramme. À la question

```
LARGEUR DU DENDROGRAMME EN CARACTERES (MINIMUM 10, MAXIMUM 279)
```

il faut répondre par un nombre entier compris entre 10 et 279, selon la largeur de l'écran ou du papier disponible pour l'impression. Notez que la largeur demandée ne concerne que le dendrogramme lui-même; à cela il faut ajouter 12 caractères à gauche pour les noms d'objets et la marge du dendrogramme, et 10 caractères à droite pour les niveaux de fusion (voir l'exemple ci-dessous).

Exemple

L'exemple ci-dessous, réalisé sous CMS, est le résultat d'un groupement à liaison proportionnelle avec *Co* = 0.5, pour les 5 mares déjà utilisées pour illustrer les statistiques de groupement, à la section du programme GROUPEMENTS. La corrélation cophénétique (*r* de Pearson) est de 0.94680. À gauche du dendrogramme se trouvent les noms des objets. À défaut de noms, le programme de groupement leur attribuera les numéros de 1 à *n*. Chaque niveau de fusion (exprimé en <u>distances</u>), indiqué à droite, correspond au trait vertical qui **commence** à sa gauche et se dirige vers le bas. Ainsi, le trait vertical identifié par la flèche a la valeur de D = 0.40000, indiquée à droite.

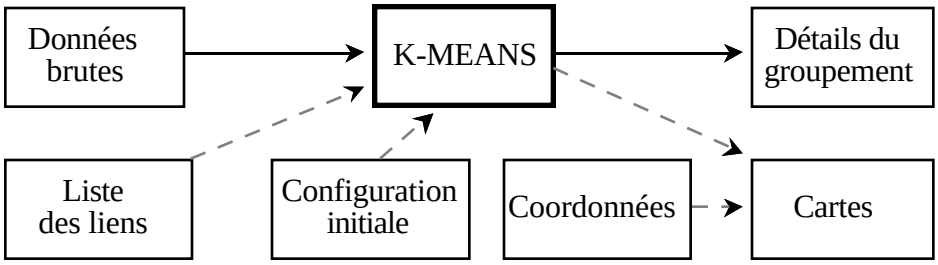
```
P R O G R A M M E   D E N D R O

Logiciel R, Version 3.0b

NOMBRE D OBJETS :    5                    [Bloc d'informations concernant
NOMBRE DE VARIABLES:    8                 la matrice de similarités utilisée]
TITRE: 5 mares de Legendre & Chodorowski (1977)
DATE 03/03/91
FONCTION  s20

D E N D R O G R A M M E
                                                     NIVEAU(D)
--------------------------------------------------------------


MARE 214   |<-----------------                         0.40000
                            |<----------------------   0.93700
MARE 212   |<-----------------                     |
                                                   |<
MARE 432   |<--------------------                  |   0.50000
                                |<--------         |   0.70000
MARE 431   |<--------------------         |        |
                                          |<---------
MARE 233   |<------------------------------
```

*K-MEANS*Macintosh *ou KMEANS*CMS/VMS

Que fait K-MEANS ?

Ce programme réalise le groupement non-hiérarchique par minimisation de la variance intragroupe, selon différentes variantes de la méthode proposée d'abord par MacQueen (1967), à laquelle celui-ci a donné le nom de méthode *k-means*. Il s'agit d'une méthode de partition d'un groupe d'objets, et non d'une méthode de classification hiérarchique. L'usager précise quel est le nombre, *k*, de groupes qu'il désire obtenir au terme du groupement. L'algorithme *k*-means suivi ici est celui décrit à la page 163 de Anderberg (1973). Ce programme réalise le groupement avec ou sans contrainte de contiguïté (spatiale ou temporelle), selon les désirs de l'usager. Il complémente donc les programmes de groupement hiérarchique du progiciel "R", programmes qui mettent en oeuvre divers algorithmes de groupement sans contrainte (GROUPEMENTS pour la version Macintosh, INTERLNK et LANCE pour les versions CMS et VMS) ou avec contrainte de contiguïté (BIOGÉO et CHRONO).

Fichiers d'entrée et de sortie



Les flèches en tirets indiquent des fichiers optionnels.

(1) Fichier de données brutes

Contrairement aux autres programmes de groupement de ce progiciel qui requièrent comme fichier d'entrée une matrice de similarités, les données sont fournies au programme K-MEANS sous la forme d'une matrice rectangulaire (p x n) de données brutes où les lignes sont les objets et les colonnes sont les variables, le tout sans identificateurs de lignes ou de colonnes. Par exemple:

23.4	12.4	3.2	77	22.6
12.6	13.2	4.9	44.1	23.6
33.4	11.8	5.5	55.3	21
45.1	12	3.1	109	22.8
50.7	11.7	4.6	67.9	23.5

D'autres algorithmes de type *k-means* utilisent plutôt une matrice de distances comme point de départ des calculs. Ce programme ne sait pas traiter les absences d'information. Celles-ci devront être comblées par l'une ou l'autre méthode d'interpolation avant le groupement; une autre méthode consiste à supprimer les objets (lignes) porteurs d'informations absentes.

N.B. Le programme minimise en fait la somme des carrés des distances *euclidiennes* des objets au centroïde de leur groupe respectif. Si on désire appliquer la méthode à des données pour lesquelles la distance euclidienne est jugée inappropriée (par exemple des données d'abondance d'espèces contenant beaucoup de zéro), on peut procéder selon les étapes suivantes (voir aussi l'exemple):
1) Calculer la matrice de similarités ou de distances de son choix à l'aide de SIMIL.
2) Réaliser une analyse en coordonnées principales.

2) Réaliser une analyse en coordonnées principales.
3) Demander au programme PCOORD d'inscrire sur un fichier de sortie un certain nombre de coordonnées principales (en général, 10 ou 15 coordonnées principales suffisent à rendre compte de presque toute la variabilité).
4) Ces coordonnées principales peuvent maintenant être fournies au programme K-MEANS comme nouvelles données brutes.

(2) Fichier de liens (optionnel)

Si l'usager décide de réaliser un groupement avec contrainte de contiguïté spatiale, il faudra en plus du fichier de données brutes fournir au programme un fichier de LIENS, comme dans BIOGEO. Voir l'exemple dans la description de ce programme. Le fichier de liens, qui peut avoir été produit par les programmes AUTOCOR (version CMS/VMS) ou CONNEXIONS (version Macintosh), aurait par exemple l'apparence suivante pour une grille de 12 points disposés en 3 lignes et 4 colonnes, mouvement de la tour; chaque paire de numéros représente un lien entre deux objets:

```
 1  2     2  3     3  4     5  6     6  7     7  8     9 10    10 11
11 12     5  1     6  2     7  3     8  4     9  5    10  6    11  7
12  8
```

(3) Fichier de coordonnées spatiales (X, Y) (optionnel)

Si on désire demander au programme de tracer la carte correspondant à chaque groupement, ce qui représente une option du programme, il faut lui fournir un fichier contenant les coordonnées des points. Ces coordonnées ne serviront qu'à pointer les objets sur la carte. Les coordonnées sont fournies en format lisible (ASCII, non en binaire) sous la forme d'entiers ou de nombres réels en degrés décimaux. Les coordonnées **ne doivent pas** être en degrés-minutes-secondes. Le nombre de paires de coordonnées doit correspondre au nombre d'objets à partitionner. Pour les versions CMS et VMS, il faut écrire un zéro avant le point décimal (par exemple, "0.376" plutôt que ".376").

Pour certaines représentations didactiques, on pourra fournir dans ce fichier des coordonnées qui ne correspondent pas exactement aux positions géographiques. Par exemple, pour analyser en une seule fois des échantillonnages répétés d'un même territoire au cours du temps, on pourra prévoir la position des objets de l'étude de façon à ce que chaque tranche de temps forme une partie séparée de l'image finale. Les coordonnées fournies dans ce fichier ne servent qu'à l'illustration; les relations spatiales ou spatio-temporelles qui sont tenues en compte lors du groupement sont uniquement celles que contient le fichier de liens.

(4) Fichier de configuration initiale (optionnel)

Seul le fichier de données brutes est nécessaire si un groupement sans contrainte est désiré. Dans bien des cas, cependant, un fichier contenant une ou plusieurs configurations initiales possibles des objets sera inclus pour améliorer la performance de l'algorithme et éviter de se trouver coincé dans un minimum local de la fonction objective (voir options 1b et 2b); on utilisera également cette option lorsqu'on utilise le programme K-MEANS pour améliorer un groupement obtenu de façon agglomérative. Ce fichier se présente sous la forme de la liste des objets que l'on attribue à chaque groupe. ATTENTION: la liste de chaque groupe doit se terminer par un zéro. Si on inscrit plusieurs configurations initiales dans ce fichier, celles-ci seront traitées à tour de rôle par le programme. Par exemple, si on désire tester deux configurations initiales d'un problème comportant quatre groupes et un total de 13 objets, le fichier pourrait se présenter comme suit:

```
1 7  3  12  0
8  2  0
10  13  4  5  0
6  9  1  0                    [fin de la première configuration initiale]
```

```
2  4  10  0
9  1  3  13  0
12  5  6  0
7  8  1  0
```
[ceci complète la seconde configuration initiale]

Si un objet n'a été assigné à aucun groupe, le programme demandera en mode conversationnel à quel groupe l'usager désire assigner l'objet en question; si on a assigné par erreur un objet à plusieurs groupes, c'est la dernière assignation qui est retenue.

(5) Fichier de résultats

Les résultats sont présentés sous la forme d'une liste d'objets membres de chaque groupe. Dans les versions CMS et VMS, les résultats du groupement avec contrainte ne sont pas présentés sous la forme de cartes; ils le sont dans la version Macintosh. Le programme peut ne fournir que les configurations initiale et finale, ou encore toutes les étapes intermédiaires. Outre la liste des membres de chaque groupe, il indique pour chaque groupe la valeur de la statistique (E) de somme des carrés des distances au centroïde, ainsi que la valeur de la statistique D (qui est la somme des valeurs de E) pour l'ensemble de la solution. Voir la section des résultats pour plus de détails sur le fichier de sortie.

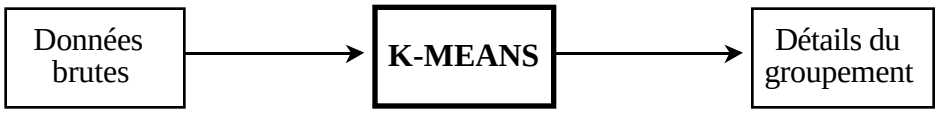
Les options du programme

La difficulté de cette méthode consiste à établir une configuration initiale des objets, c'est-à-dire une division initiale des objets en *k* groupes, configuration qui soit suffisamment près de la solution minimisant la somme des variances intra-groupes pour permettre à l'algorithme de converger vers celle-ci. Les solutions à ce problème qui sont incluses dans le programme sont les suivantes.

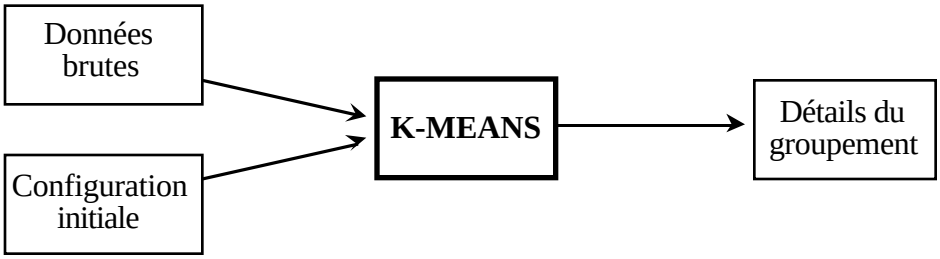
(1) Groupement sans contrainte

Pour les groupements sans contrainte de contiguïté, trois options sont disponibles.

1a) La méthode dite "de Stony Brook", car employée par R. R. Sokal à cette université. Elle consiste à réaliser *N* itérations, chacune démarrant à partir d'une distribution initiale aléatoire des objets dans les *k* groupes. À chaque itération, on calcule une statistique D et la solution qui minimise D est retenue comme solution finale. D est la somme, pour tous les groupes, des sommes des carrés des distances des membres du groupe à leur centroïde (Späth, 1980, p. 73).



1b) L'usager peut fournir sa propre configuration initiale. La manière de le faire est décrite plus haut. Une configuration initiale présumément proche de la solution optimale peut avoir été obtenue d'une autre méthode de groupement ou d'ordination; ceci est certainement la méthode la plus rapide et la plus efficace pour éviter de se trouver coincé dans un minimum local de la fonction D. Dans d'autres cas, la solution initiale à tester peut être connue par hypothèse.

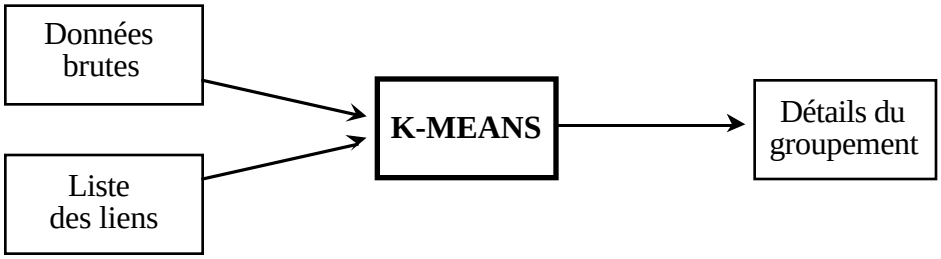


1c) La méthode MODULO (Späth, 1980, p. 67) qui consiste à établir la configuration initiale en attribuant l'objet 1 au groupe 1, l'objet 2 au groupe 2, ... , l'objet *k* au groupe *k*, l'objet *k* +1 au groupe 1, etc.

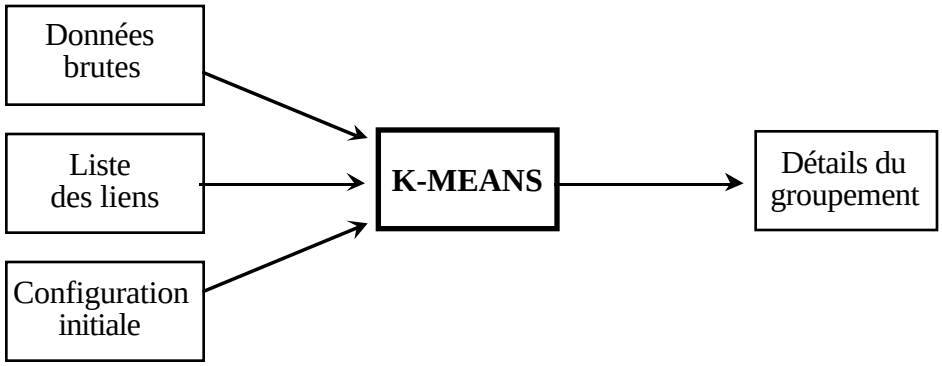
(2) Groupement avec contrainte

Pour les groupements avec contrainte de contiguïté, si la contrainte est uni-dimensionnelle (séquence temporelle d'échantillonnage ou transect), il suffit de le préciser en réponse à l'une des questions du programme et l'algorithme assumera que dans la liste, les objets successifs sont adjacents. Si au contraire les objets sont répartis dans un espace à deux dimensions ou plus, on fournit les contraintes sous la forme d'un fichier de LIENS, comme dans BIOGEO. Les deux solutions suivantes sont disponibles pour établir la configuration initiale.

2a) La méthode de Stony Brook, comme dans (1a) ci-dessus.



2b) Votre propre configuration initiale, connue de par la théorie ou encore obtenue d'une autre méthode de groupement ou d'ordination, comme dans (1b) ci-dessus. À cause de la nature même de son algorithme (minimisation de la variance intragroupe), ce programme peut être utile pour préciser la position des frontières entre les groupes établis par le programme de groupement agglomératif sous contrainte BIOGEO.



Exemples

Les deux exemples ci-dessous illustrent l'utilisation du programme pour calculer un groupement sous contrainte de contiguïté spatiale. Dans le premier cas, le groupement sera réalisé à partir de 10 configurations au hasard; voir (1) [les numéros se réfèrent aux numéros en marge gauche des exemples ci-dessous]. Dans le second cas (2), un fichier comportant deux configurations initiales sera fourni au programme. Le programme d'appel, dont le dialogue est présenté ci-dessous (exemples réalisés sous CMS), demande le nom des divers fichiers à tour de rôle; les réponses de l'usager sont soulignées et en gras. Les questions posées par la version Macintosh sont essentiellement les mêmes, quoique leur formulation pourra parfois différer légèrement.

Les données analysées ci-dessous sont les mêmes qui ont servi d'exemple pour illustrer le programme BIOGÉO. Puisque le programme K-MEANS travaille à partir d'un fichier de données rectangulaire (objets x variables), la matrice de similarités a été soumise au préalable au programme d'analyse en coordonnées principales PCOORD, qui a calculé les coordonnées de chaque objet dans un espace euclidien (voir la note à la fin de la section portant sur le fichier de données brutes); les deux premières coordonnées principales ont été retenues, car elles seules correspondaient à des valeurs propres positives. Puisque l'analyse en coordonnées principales fournit une représentation euclidienne du nuage de points et puisque l'on désire justement que le groupement minimise la somme des carrés de ces mêmes distances euclidiennes au centroïde des différents groupes, on demandera au programme (3) de ne pas faire de transformation des données. Si on avait eu affaire à une série de variables non commensurables, on aurait alors demandé au programme d'effectuer d'abord un centrage et une réduction des données, suivant la même logique qu'en analyse des composantes principales.

Exemple 1: à partir de 10 configurations au hasard

```
Kmeans
ATTENTION! Ce programme ne traite pas l'absence d'information.
Quel est le nom du fichier des donnees brutes (lignes = objets,
colonnes = variables)? (Par defaut: "... data a")
fichier  pcoord  a

Quel est le nom du fichier des LIENS entre localites
(s'il y a lieu)? (Par defaut: "... data a")
fichier  liens  a

Quel est le nom du fichier contenant la ou les CONFIGURATION(s)
INITIALE(s), si vous desirez en fournir? (Par defaut: "... data a")

Quel est le nom du fichier ou les RESULTATS devront etre ecrits?
(Optionnel; par defaut: "RESKM OUT a")
fichier  res1  a

Execution begins...
 P R O G R A M M E   K - M E A N S  avec contraintes

 Auteur: Alain Vaudor

 Nombre d'objets
57
 Nombre de variables
2
 Nombre de groupes
4
 Type de groupement:
```

```
      Type de groupement:
        0: Groupement sans contrainte
        1: Groupement avec contrainte de contiguite en 1 dimension
        2: Groupement avec contraintes generales (fichier de liens obligatoire)
     2
      Options:
        1: Au hasard (methode de Stony Brook)
        2: Votre fichier de configuration(s) initiale(s)
(1)  1
      Nombre d'essais ?
(1)  10
      Options:
        1: Imprimer tous les resultats intermediaires
        2: Configurations initiale et finale seulement
     2
      Options:
        0: Pas de transformation des donnees
        1: Transformation en variables centrees reduites
(3)  0
      Tapez un chiffre (petit entier) pour indiquer le
      point de depart du generateur de nombres aleatoires
     5
      Fin du programme.
```

Exemple 2: à partir d'un fichier comportant deux configurations initiales

```
     Kmeans
     ATTENTION! Ce programme ne traite pas l'absence d'information.
     Quel est le nom du fichier des donnees brutes (lignes = objets,
     colonnes = variables)? (Par defaut: "... data a")
     fichier  pcoord  a

     Quel est le nom du fichier des LIENS entre localites
     (s'il y a lieu)? (Par defaut: "... data a")
     fichier  liens  a

     Quel est le nom du fichier contenant la ou les CONFIGURATION(s)
     INITIALE(s), si vous desirez en fournir? (Par defaut: "... data a")
(2)  fichier  init  a

     Quel est le nom du fichier ou les RESULTATS devront etre ecrits?
     (Optionnel; par defaut: "RESKM OUT a")
     fichier  res2  a

     Execution begins...
      P R O G R A M M E   K - M E A N S  avec contraintes

      Auteur: Alain Vaudor

      Nombre d'objets
     57
      Nombre de variables
     2
      Nombre de groupes
     4
```

```
      Type de groupement:
        0: Groupement sans contrainte
        1: Groupement avec contrainte de contiguite en 1 dimension
        2: Groupement avec contraintes generales (fichier de liens obligatoire)
     2
      Options:
        1: Au hasard (methode de Stony Brook)
        2: Votre fichier de configuration(s) initiale(s)
(2)  2
      Nombre d'essais ?
     2
      Options:
        1: Imprimer tous les resultats intermediaires
        2: Configurations initiale et finale seulement
     2
      Options:
        0: Pas de transformation des donnees
        1: Transformation en variables centrees reduites
(3)  0
      Fin du programme.
```

Contenu du fichier de résultats, exemple 2

Les sorties présentées ci-dessous ont été produites par la version Macintosh du programme. Les versions CMS et VMS produisent un fichier de sortie ASCII identique, mais sont incapables de tracer les cartes.

Première configuration initiale:

La première configuration initiale soumise au programme divisait les objets en quatre groupes, comme suit (la fin de chaque groupe est marquée par un zéro):

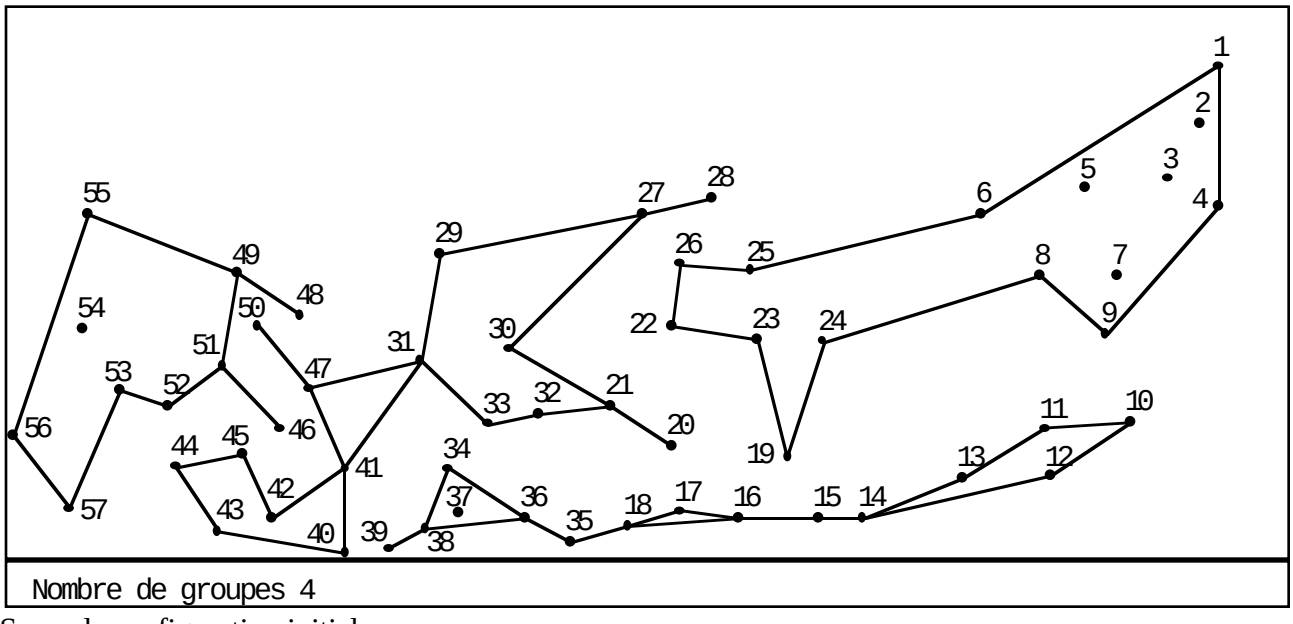
```
42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57  0
22 23 24 25 26 27 28 29 30 31  0
14 15 16 17 18 19 20 21 32 33 34 35 36 37 38 39 40 41  0
 1  2  3  4  5  6  7  8  9 10 11 12 13  0
```

En se référant à la carte ci-dessous, on voit que selon cette hypothèse, les localités seraient divisées en quatre blocs d'à peu près égale importance: le premier à gauche, les deux suivants au centre (partie du haut, partie du bas) et le dernier à droite. La statistique D, qui représente la somme, pour les différents groupes, des sommes (E) des carrés des distances au centroïde, a au départ la valeur de 11.72596; en déplaçant certains objets, l'algorithme réussit à réduire cette valeur à D (ou: Somme des E) = 8.35474. On verra avec l'exemple suivant que ce résultat est encore très éloigné de la valeur optimale; cet exemple a justement été présenté pour montrer que les algorithmes de type *k-means* peuvent souvent, selon la configuration initiale qui leur est fournie, ne pas converger vers la valeur minimum de la statistique D.

```
      Essai no 1

        Etape  initiale
           Groupe  1: 42  43  44  45  46  47  48  49  50  51  52  53  54  55  56
                      57
                   E =   3.68762
           Groupe  2: 22  23  24  25  26  27  28  29  30  31
```

```
   Groupe  2: 22  23  24  25  26  27  28  29  30  31
           E =   1.75314
   Groupe  3: 14  15  16  17  18  19  20  21  32  33  34  35  36  37  38
              39  40  41
           E =   3.84957
   Groupe  4:  1   2   3   4   5   6   7   8   9  10  11  12  13
           E =   2.43562
   Somme E =       11.72596
Etape 1
   Groupe  1: 40  43  44  46  48  49  51  52  53  54  55  56  57
           E =   2.75096
   Groupe  2: 19  20  21  22  23  27  28  29  30  31  32  33  41  42  45
              47  50
           E =   2.35024
   Groupe  3: 12  14  15  16  17  18  34  35  36  37  38  39
           E =   2.48375
   Groupe  4:  1   2   3   4   5   6   7   8   9  10  11  13  24  25  26
           E =   2.44078
   Somme E =       10.02573
Etape 2
   Groupe  1: 46  48  49  51  52  53  54  55  56  57
           E =   1.48061
   Groupe  2: 20  21  27  28  29  30  31  32  33  40  41  42  43  44  45
              47  50
           E =   2.21594
   Groupe  3: 10  11  12  13  14  15  16  17  18  34  35  36  37  38  39
           E =   2.82830
   Groupe  4:  1   2   3   4   5   6   7   8   9  19  22  23  24  25  26
           E =   1.82989
   Somme E =        8.35474
```

La carte produite par le programme est la suivante. Les groupes sont <u>entourés</u> d'une enveloppe; les points individuels qui apparaissent à l'intérieur d'une enveloppe, par exemple 2, 3, 5 et 7, sont membres du même groupe que 1, 4, 6, etc.



<u>Seconde configuration initiale</u>:

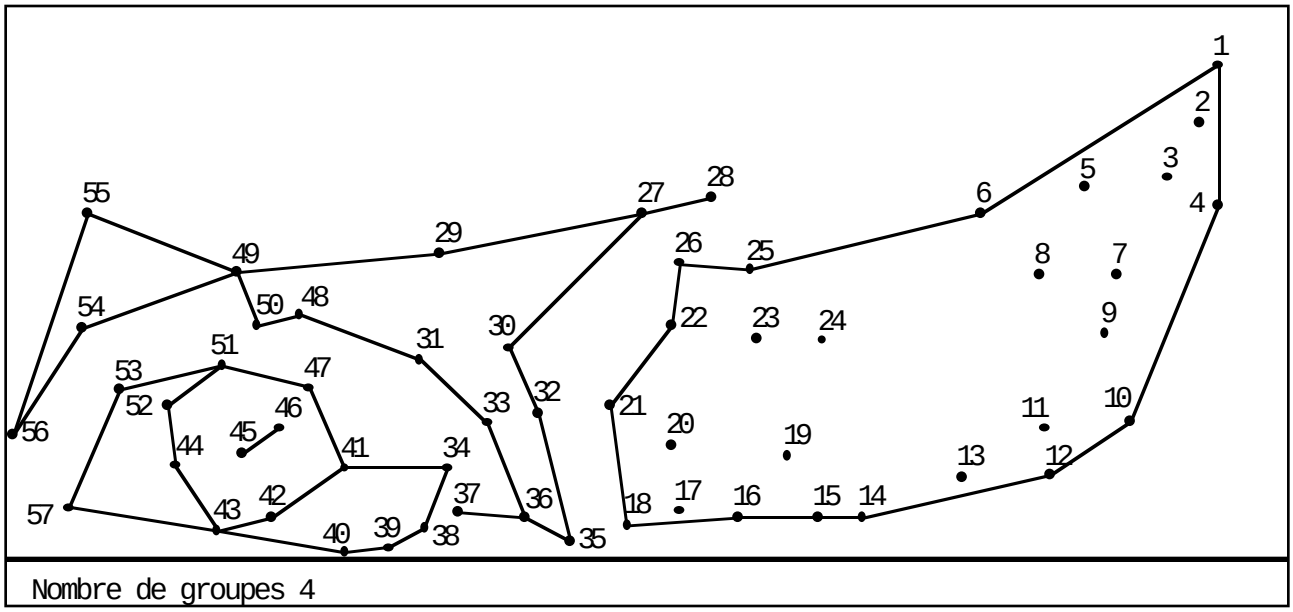
La seconde configuration initiale fournie au programme K-MEANS est la solution à quatre groupes qu'avait produite BIOGÉO; voir la section "Contenu du fichier de résultats" de ce programme:

```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26  0
27 28 29 30 31 32 33 35 36 37 48 49 50 54 55 56  0
45 46  0
34 38 39 40 41 42 43 44 47 51 52 53 57  0
```

Cette configuration initiale produit une statistique D (ou: Somme des E) = 7.71485 que le programme n'a pas réussi à améliorer par interchange d'objets entre les groupes.

```
Essai no 2

  Etape  initiale
     Groupe  1:  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
                16  17  18  19  20  21  22  23  24  25  26
             E =   4.48868
     Groupe  2: 27  28  29  30  31  32  33  35  36  37  48  49  50  54  55
                56
             E =   0.87511
     Groupe  3: 45  46
             E =   0.17951
     Groupe  4: 34  38  39  40  41  42  43  44  47  51  52  53  57
             E =   2.17156
     Somme E =        7.71485
  La solution initiale n'a pu être améliorée.
```

La carte produite par le programme est la suivante. Elle montre les mêmes assignations de groupes que la carte présentée dans l'exemple du programme BIOGÉO, à l'étape "4 groupes". On remarque que le groupe (45, 46) est à l'intérieur du "beigne" formé par les deux enveloppes (extérieure et intérieure) dessinées pour le groupe (34 à 57). **N.B.** — Pour augmenter la résolution, on peut agrandir n'importe quelle partie de l'image en l'entourant d'un cadre à l'aide de la souris.



LANCECMS/VMS

Que fait LANCE ?

Ce programme réalise le groupement agglomératif selon l'algorithme proposé par Lance & Williams (1966, 1967). Les méthodes disponibles incluent les groupements à liens simples, à liens complets, selon l'association moyenne (UPGMA), à poids proportionnels (WPGMA), centroïde (UPGMC), médian (WPGMC), ainsi que la famille de méthodes connues sous le vocable de groupement flexible. La méthode de Ward (1963), ou méthode de minimisation de la variance intra-groupe, a récemment été introduite dans ce même programme, suivant en cela la proposition de Everitt (1980). Pour les méthodes autres que liens simples, liens complets, UPGMA, UPGMC, WPGMA, WPGMC et Ward, le programme demande à l'usager de fournir les quatre paramètres *alpha*$_j$, *alpha*$_m$, *bêta* et *gamma* qui déterminent la stratégie de groupement que réalisera l'algorithme de Lance & Williams. Le tableau 2 décrit les valeurs de ces paramètres dans les différents cas. On consultera les ouvrages en référence dans l'en-tête du tableau pour connaître le rôle de ces paramètres dans la stratégie de groupement.

Tableau 2 — Valeurs des paramètres *alpha*$_j$, *alpha*$_m$, *bêta* et *gamma* de l'équation générale de Lance & Williams (1966), pour les différents types combinatoires de groupement séquentiel agglomératif. Inspiré de Sneath & Sokal (1973), Legendre & Legendre (1984a) et Jain & Dubes (1988).

Méthode de groupement	Paramètres du modèle combinatoire α_j	α_m	β	γ
Liens simples	0.5	0.5	0	-0.5
Liens complets	0.5	0.5	0	0.5
Groupements moyens:				
Association moyenne (UPGMA)	$n_j/(n_j+n_m)$	$n_m/(n_j+n_m)$	0	0
Poids proportionnels (WPGMA)	0.5	0.5	0	0
Centroïde (UPGMC)	$n_j/(n_j+n_m)$	$n_m/(n_j+n_m)$	$-\alpha_j\alpha_m$	0
Médian (WPGMC)	0.5	0.5	-0.25	0
Groupement flexible	$[\alpha_j + \alpha_m + \beta = 1; \quad \alpha_j = \alpha_m; \quad -1 \le \beta \le 1]$			0
Méthode de Ward	$(n_j+n_g)/(n_j+n_m+n_g)$	$(n_m+n_g)/(n_j+n_m+n_g)$	$-n_g/(n_j+n_m+n_g)$	0

Le programme d'appel LANCE met en route à tour de rôle trois programmes différents: (1) un programme de tri qui récrit la matrice de similarités en ordre de similarités décroissantes (nécessaire pour les tests *a posteriori* du groupement), (2) le programme de groupement lui-même, et finalement (3) le programme qui trace le dendrogramme. L'usager peut demander à ce dernier programme de calculer différentes statistiques (chaîne des liens primaires, corrélations cophénétiques, distance de Gower, coefficients d'efficacité) qui ont été décrites à la fin de la section portant sur le programme GROUPEMENTS. Le programme LANCE n'existe que dans les versions CMS et VMS de "R".

Fichiers d'entrée et de sortie

(1) Le fichier d'entrée

Le fichier d'entrée peut contenir une matrice de similarités ou de distances, écrit par SIMIL ou IMPORT (puisque le programme LANCE n'existe qu'en versions CMS et VMS).

Le nombre maximum d'objets qui peuvent être traités par ce programme est fixé par le paramètre MAXNOBJ, en début du programme. Ce paramètre peut être ajusté à la taille des problèmes à traiter, avant la compilation.

(2) Le fichier de sortie

Ce fichier contient le dendrogramme décrivant le groupement agglomératif, ainsi que les statistiques de groupement. Celles-ci sont décrites en détail à la section relative au programme GROUPEMENTS. Si on a attribué des noms aux objets dans le fichier de données brutes soumis à SIMIL (10 premiers caractères), le dendrogramme présente ces identificateurs, au lieu des numéros d'ordre que le programme attribuera autrement aux objets.

Les questions du programme

Après que le programme d'appel ait demandé le nom des fichiers d'entrée et de sortie, le programme de groupement lui-même demande à l'usager de choisir la méthode de groupement qu'il désire employer. Si le choix se porte sur l'option 6, il aura encore à fournir les paramètres $alpha_j$, $alpha_m$, *bêta* et *gamma* dont le programme a besoin pour réaliser les calculs. Voir le tableau 2.

```
DESIREZ-VOUS
     1- Association moyenne  (UPGMA)
     2- Poids proportionnels (WPGMA)
     3- Groupement centroïde (UPGMC)
     4- Groupement médian    (WPGMC)
     5- Méthode de Ward (1963)
     6- Autres groupements combinatoires
```

Ce programme est le seul parmi les programmes de groupement de "R" à permettre l'emploi d'une matrice de distances aussi bien que d'une matrice de similarités. La question suivante est posée pour déterminer de quel type de matrice il s'agit:

```
LA MATRICE D'ENTREE EST DE SIMILARITES OU DE DISTANCES? (S ou D)
```

À cette question, on doit répondre par une lettre: *S* ou *s* s'il s'agit d'une matrice de similarités, *D* ou *d* pour une matrice de distances. Il reste cependant que le mode normal de fonctionnement de ce programme demande une matrice de similarités, si bien que l'ajustement des calculs ne s'étend pas jusqu'aux statistiques associées au dendrogramme, qui est calculé par un programme différent. L'avertissement suivant est fourni à l'usager qui aura choisi de traiter une matrice de distances:

```
Lecture de la matrice de DISTANCES.
```

```
Lecture de la matrice de DISTANCES.
Les tests a posteriori ne sont tous corrects que
pour les matrices de SIMILARITE. Si vous les demandez,
 - la chaîne des liens primaires sera donc erronée;
 - le signe des corrélations cophénétiques sera inversé;
 - la distance de Gower sera incorrecte;
 - les mesures d'entropie seront correctes.
```

Les questions suivantes sont posées par le programme DENDRO qui se charge de tracer le dendrogramme et de calculer les statistiques de groupement; voir la description de ces statistiques à la section portant sur le programme GROUPEMENTS. La largeur du dendrogramme qui sera dessiné est fixée par l'utilisateur, qui doit déterminer le nombre de caractères d'imprimerie qui seront utilisés pour tracer le dendrogramme. À la question

```
LARGEUR DU DENDROGRAMME EN CARACTERES (MINIMUM 10, MAXIMUM 279)
```

il faut répondre par un nombre entier compris entre 10 et 279, selon la largeur de l'écran ou du papier disponible pour l'impression. Notez que la largeur demandée ne concerne que le dendrogramme lui-même; à cela il faut ajouter 12 caractères à gauche pour les noms d'objets et la marge du dendrogramme, et 10 caractères à droite pour les niveaux de fusion (voir l'exemple ci-dessous).

Note sur la méthode de Ward

La méthode de minimisation de la variance de Ward (1963) fusionne les objets ou les groupes de façon à minimiser la somme des carrés des distances au centroïde de chaque groupe. Les calculs par l'algorithme général de Lance & Williams se font sur la matrice des **distances au carré** D^2. Les dendrogrammes peuvent être représentés de différentes façons, selon les auteurs. Ainsi, Jain & Dubes (1988) utilisent directement comme échelle horizontale du dendrogramme les niveaux de fusion obtenus de l'algorithme de groupement, exprimés en distances au carré. Everitt (1980) utilise plutôt une statistique de somme, pour les différents groupes, des sommes des carrés des écarts au centroïde de chaque groupe, ou E.S.S. dans le livre de Everitt, qui peut également se calculer comme la somme, pour les différents groupes k, des valeurs $e_k^2 = \Sigma(D^2)/n_k$. L'échelle de Jain & Dubes est une simple transformation linéaire de l'échelle utilisée par Everitt. Dans le manuel SAS (1985), enfin, on recommande d'employer l'une ou l'autre des statistiques suivantes comme échelle du groupement: soit la statistique E.S.S. de Everitt divisée par la somme totale des carrés des écarts, ce qui produit des proportions de variance (manuel SAS 1985, p. 267), soit encore le *R-carré semipartiel* qui se calcule comme la somme des carrés des écarts inter-groupes divisée par la somme totale des carrés des écarts (manuel SAS 1985, p. 272 et 281); il s'agit de nouveau de transformations linéaires des échelles de Everitt et de Jain & Dubes.

Notez que toutes les mesures ci-dessus sont essentiellement des distances au carré. Dans le programme LANCE, nous employons plutôt la **racine carrée** des distances de fusion au carré produites par l'algorithme combinatoire de Lance & Williams et utilisées par Jain & Dubes. Il y a à cela deux avantages. D'une part, les dendrogrammes ont une apparence mieux balancée que ceux produits par les méthodes énumérées ci-dessus. D'autre part, c'est la distance la plus appropriée lorsqu'on veut comparer, par corrélation de matrices ou par la distance de Gower, la matrice cophénétique du dendrogramme à la matrice des distances ou des similarités d'origine.

Exemple

L'exemple ci-dessous, réalisé sous CMS, est le résultat d'un groupement selon la méthode de Ward, pour les 5 mares déjà utilisées pour illustrer les statistiques de groupement, à la section du programme GROUPEMENTS. Quatre nouvelles variables ont été obtenues d'une analyse en

coordonnées principales de la matrice de similarités (coefficient S20) calculée à partir des données d'origine; la distance euclidienne (D01) a ensuite été calculée entre les objets (mares) pour ces nouvelles variables, ce afin d'illustrer les réponses du programme lorsque le fichier d'entrée contient une matrice de distances.

```
Quel est le nom du fichier contenant la matrice de SIMILARITES
de type SIMIL? (Par defaut: "... data a")
mares  d1  a

Quel nom doit recevoir le fichier contenant le dendrogramme
et les tests? (Par defaut: "RESULTAT listing a")
mares  dendr-d1  a

Execution begins...              Annonce le début de l'exécution du programme de tri
Execution begins...      Annonce le début de l'exécution du programme de groupement

 P R O G R A M M E   L A N C E -- Modele general de groupement agglomeratif.

 Version 2.2b (Modifie pour SIMIL 3.0 / Inclut Ward)
 Auteur: A. VAUDOR

 DESIREZ-VOUS
      1- Association moyenne  (UPGMA)
      2- Poids proportionnels (WPGMA)
      3- Groupement centroide (UPGMC)
      4- Groupement median    (WPGMC)
      5- Methode de Ward (1963)
      6- Autres groupements combinatoires
5

 LA MATRICE D'ENTREE EST DE SIMILARITES OU DE DISTANCES? (S ou D)
d
 Lecture de la matrice de DISTANCES.
 Les tests a posteriori ne sont tous corrects que
 pour les matrices de SIMILARITE. Si vous les demandez,
  - la chaine des liens primaires sera donc erronee;
  - le signe des correlations cophenetiques sera inverse;
  - la distance de Gower sera incorrecte;
  - les mesures d'entropie seront correctes.

 Fin du groupement.

Execution begins...      Début de l'exécution du programme qui trace le dendrogramme

 P R O G R A M M E   D E N D R O

 Dendrogramme, chaine des liens primaires, tests entre les groupes
 Version 3.0b
 AUTEUR: A. VAUDOR

 VOULEZ-VOUS LA CHAINE DES LIENS PRIMAIRES ( o ou n) ?
o
 VOULEZ-VOUS LES TESTS A POSTERIORI:
 CORRELATIONS COPHENETIQUES, DISTANCE DE GOWER ET ENTROPIE?
```

```
n
 LARGEUR DU DENDROGRAMME EN CARACTERES ( MINIMUM:  10 MAXIMUM 279)
50
 Fin du programme.
```

Contenu du fichier de résultats

À gauche du dendrogramme se trouvent les noms des objets. À défaut de noms, le programme de groupement leur aurait attribué les numéros de 1 à *n*. Chaque niveau de fusion (exprimé en distances), indiqué à droite, correspond au trait vertical qui **commence** à sa gauche et se dirige vers le bas. Ainsi, le trait vertical identifié par la flèche a la valeur de D = 0.50000, indiquée à droite.

```
P R O G R A M M E   D E N D R O

Logiciel R, Version 3.0b

NOMBRE D OBJETS :    5
NOMBRE DE VARIABLES:    4
```

[Bloc d'informations concernant la matrice de similarités utilisée]

```
TITRE: 5 mares de Legendre & Chodorowski (1977)
DATE 03/03/91
FONCTION  d01

D E N D R O G R A M M E
                                                               NIVEAU(D)
------------------------------------------------------------------------

MARE 431   |<------------------                                  0.50000
                              |<------------                     0.81854
MARE 432   |<------------------           |
                                          |<-----------------   1.29596
MARE 233   |<-------------------------------                 |
                                                              |<
MARE 212   |<--------------                                   |   0.40000
                          |<----------------------------------
MARE 214   |<--------------
```

MANTEL

Que fait MANTEL ?

Ce programme calcule la statistique *Z* de Mantel (1967) entre deux matrices de similarité ou de distance, ainsi que les formes dérivées décrites ci-dessous: tests de Mantel partiels, corrélogramme de Mantel. La signification de la statistique de Mantel peut être évaluée par permutations, ou encore par l'approximation normale décrite par Mantel (statistique appelée *t* par Mantel, dont la distribution est asymptotiquement normale). Comme la probabilité obtenue du test approximatif converge rapidement vers la probabilité obtenue par permutations, il devient inutile d'employer le test par permutations là où il serait le plus coûteux en temps machine, soit les problèmes comportant de nombreux objets. Legendre & Fortin (1989) présentent un bref exposé de cette méthode de test par permutations.

Notez que la statistique *Z* fournie par ce programme a comme valeur la moitié de celle de Mantel (1967), car les calculs sont effectués sur des demi-matrices symétriques de similarité ou de distance; cependant, la statistique *t* de Mantel de même que la valeur standardisée de *Z* selon Hubert sont calculées comme si les matrices étaient carrées. La statistique de Mantel centrée réduite (*r*) n'est pas affectée par le calcul sur la demi-matrice.

Outre ses applications à l'analyse spatiale, le test de Mantel est utilisé dans une foule d'autres situations statistiques. Hubert *et al.* (1982), de même que McCune & Allen (1985), Burgman (1987), Hudon & Lamarche (1989) et Legendre & Fortin (1989) ont testé la conformité de modèles à des données par la méthode du test de Mantel. Legendre & Troussellier (1988) de même que Legendre & Fortin (1989) ont employé les tests de Mantel partiels de Smouse-Long-Sokal pour une modélisation de type causal. Sokal *et al.* (1987) ont proposé de limiter les permutations lors du test de Mantel de façon à évaluer laquelle de deux hypothèses concurrentes (H_1) est la plus conforme aux données; un exemple est fourni à la section portant sur les permutations limitées (Options du programme, section 8, ci-dessous).

Fichiers d'entrée et de sortie

Les questions posées par le programme EXEC à propos des fichiers d'entrée sont nombreuses et reflètent la multiplicité des options offertes par le programme. Lisez-les attentivement avant d'y répondre.

Les tests de MANTEL simples requièrent deux matrices, **A** et **B**. Les tests partiels exigent la présence d'une troisième matrice **C** en plus des fichiers **A** et **B**. Enfin, les corrélogrammes requièrent **B** et un fichier contenant les classes de distance.

(1) Fichier d'entrée B

La matrice **B** doit toujours être présente, et il doit toujours s'agir d'un fichier binaire produit par SIMIL, IMPORT-EXPORT (version Macintosh) ou IMPORT (version VMS ou CMS). Cette matrice peut aussi bien représenter une matrice de SIMILARITÉS qu'une matrice de DISTANCES.

(2) Fichier d'entrée A

La matrice **A**, quant à elle, peut prendre plusieurs formes, énumérées ci-dessous.

(2.1) Fichier binaire de similarités

La matrice **A** peut être fournie au programme sous la forme d'un fichier binaire de type SIMIL, IMPORT-EXPORT (version Macintosh) ou IMPORT (version VMS ou CMS), si l'usager le désire.

Tout comme la matrice **B**, cette matrice peut être de type SIMILARITES ou DISTANCES. Il est cependant souhaitable que les matrices **A** et **B** soient de type identique.

(2.2) Fichier de classes de distance

Dans le cas d'un corrélogramme, une série de matrices **A** seront calculées par le programme à partir des informations données dans le fichier des classes de distance (décrit ci-dessous), et utilisées à tour de rôle pour le calcul des tests de Mantel correspondant à chaque classe de distance.

Le fichier de classes de distances, utilisé pour le calcul d'un corrélogramme de Mantel, est appelé par CLASSEF dans la déclaration des noms de fichiers du programme, ainsi que dans les fichier EXEC et COM des versions pour grands ordinateurs. Ce fichier est en caractères lisibles (et non en binaire). Il contient une matrice triangulaire supérieure de classes de distance entre les objets, sans la diagonale. Pour les petits problèmes, ce fichier peut être écrit à la main par l'usager à l'aide de son éditeur ASCII. Pour les problèmes plus importants, il peut être préparé à l'aide du programme AUTOCOR (voir ce programme), ou à l'aide de tout autre programme spécifique écrit par l'usager.

Dans ce fichier, les entiers 1, 2, 3, etc. représentent les classes de distance. Un test de Mantel sera réalisé pour chaque classe de distance présente dans le fichier: le programme fabriquera une matrice **A** contenant des "1" pour toutes les paires d'objets appartenant à la classe de distance qu'il est en train de tester et des "0" pour toutes les autres paires d'objets. Si une classe “0” ou négative est présente, aucun test de Mantel ne sera réalisé pour cette classe. Le fichier suivant serait une matrice CLASSEF acceptable pour un ensemble de 6 objets:

```
1   1   2   3   3
    1   2   3   3
        2   3   3
            1   1
                1
```

(2.3) Grille régulière

La matrice **A** peut être calculée par le programme sur déclaration de la largeur de la grille rectangulaire régulière que forment les points. Le programme lit le nombre total de points dans l'en-tête du fichier contenant la matrice **B**, d'où il peut déduire la hauteur de la grille régulière.

(2.4) Fichiers de coordonnées géographiques (DMS ou décimales)

Le fichier des coordonnées à partir duquel sera calculée la matrice **A** est un fichier ASCII. Les coordonnées sont écrites en format libres. Il peut s'agir de coordonnées sur un plan cartésien, ou encore de coordonnées terrestres en degrés, minutes et secondes (DMS) ou en degrés décimaux. Par exemple, on pourra écrire 45 15 36 (en DMS), ce qui est équivalent à 45.26 (en degrés décimaux). On inscrit la latitude d'abord, la longitude ensuite. Le programme offre le choix de calculer la distance par la formule de la distance euclidienne (coordonnées planes) ou en suivant la courbure de la terre (coordonnées sur une sphère); dans ce dernier cas, les distances sont exprimées en milles marins.

(3) Fichier d'entrée C

La matrice **C** employée pour les tests de Mantel partiels est toujours une matrice binaire de similarités ou de distances de type SIMIL, IMPORT-EXPORT (version Macintosh) ou IMPORT (version VMS ou CMS), tout comme la matrice **B**.

(4) Fichier de sortie

Sur grands ordinateurs, les résultats du test de Mantel sont présentés à l'écran et non pas dans un fichier de SORTIE. L'utilisateur CMS peut utiliser la procédure CON décrite en page 2 pour

conserver les résultats sur fichier, s'il le désire. Sur Macintosh au contraire, les résultats n'apparaissent pas à l'écran mais sont plutôt inscrits sur un fichier de sortie; l'usager se voit offrir la possibilité de désigner directement l'imprimante comme son médium de sortie, ou encore de conserver les résultats dans un fichier auquel il est invité à donner un nom. Voir la section Résultats pour plus de renseignements sur la signification des résultats.

Limites du programme

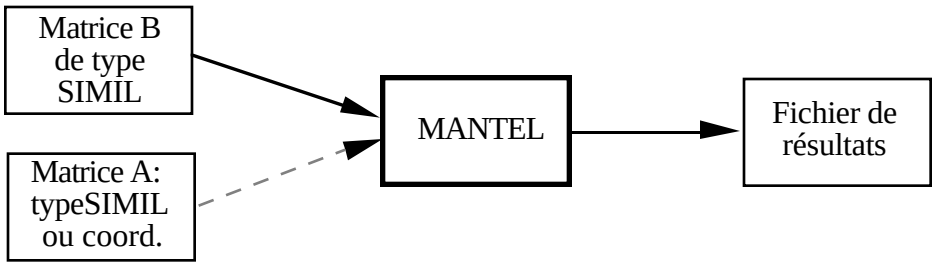
Les versions du programme pour grands ordinateurs sont limitées par deux paramètres que l'on trouvera en début de programme. Il s'agit du nombre maximum d'objets qui peuvent être traités (ex. MAXNOBJ = 1000), puis du plus grand nombre d'objets pour lesquels on autorise les tests par permutations (ex. PETITNOBJ = 200). La version Macintosh ne contient aucune de ces limites; le programme utilise de façon dynamique la mémoire disponible dans l'appareil. Un message devrait apparaître (sous FINDER) si l'ordinateur manque de mémoire vive pour réaliser le calcul. Notez que le temps de calcul augmente approximativement comme le carré du nombre d'objets.

Les options du programme

Les options disponibles dans ce programme permettent la comparaison de deux matrices, les tests de Mantel partiels de même que le calcul du corrélogramme de Mantel.

(1) Option 0: Mantel entre deux matrices

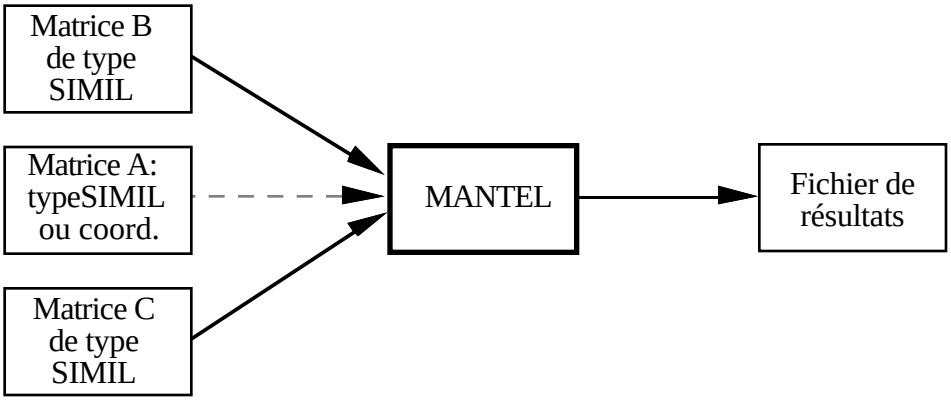
Le programme demande le nom des deux matrices (**A** et **B**) en plus du type de la matrice **A**.



Avec les options (2.3) et (2.4) relatives à la matrice **A** (grille régulière, ou fichier de coordonnées en degrés-minutes-secondes ou en degrés décimaux: voir ci-dessus), l'usager peut demander de transformer les distances calculées par le programme pour la matrice **A** en 1/D ou en $1/D^2$.

(2) Options 1 à 3: Tests de Mantel partiels

Le programme propose un choix de tests partiels. Ces différentes méthodes exigent toutes que trois matrices soient présentes (**A, B** et **C**). Encore ici, les matrices **B** et **C** sont toujours obligatoirement de type SIMIL alors que la matrice **A** peut prendre l'une ou l'autre des formes décrites ci-dessus.



Option 1: Méthode de Dow & Cheverud (1985). Statistique: **(A * (B - C))** où ***** représente la somme des produits de Mantel et permet de reconnaître les deux blocs à permuter. On peut exprimer cette statistique comme suit: $\Sigma\ [a_{ij} * (b'_{ij} - c'_{ij})]$ (statistique non centrée réduite) ou $\Sigma\ [a'_{ij} * (b'_{ij} - c'_{ij})']/(n-1)$ (statistique centrée réduite), où le signe *prime* (') représente une valeur centrée réduite.

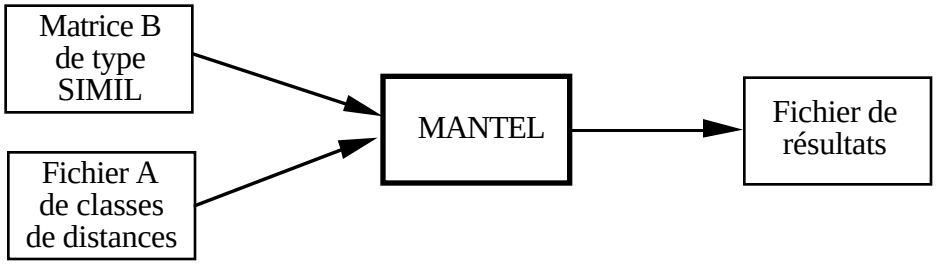
Option 2: Méthode de Smouse, Long & Sokal (1986). Statistique: **(A*B.C)**. Cette statistique est en fait la corrélation partielle entre les valeurs de **A** et de **B** conditionnellement à **C**. Le calcul est réalisé en calculant d'abord **A '** qui est la matrice des résidus de la régression des valeurs de **A** contre celles de **C**, puis **B '** qui est la matrice des résidus de la régression des valeurs de **B** contre celles de **C**, après avoir centré et réduit les valeurs au sein de chacune de ces matrices; on fait ensuite un test de Mantel entre **A '** et **B '**. Ceci n'est qu'une autre façon de calculer la corrélation partielle recherchée.

Option 3: Méthode de Hubert (1985). Statistique: **(A * (BC))**. On peut exprimer le détail de cette statistique comme suit: $\Sigma\ [a_{ij} * (b_{ij} * c_{ij})]$ (statistique non centrée réduite) ou $\Sigma\ [a'_{ij} * (b_{ij} * c_{ij})']/(n-1)$ (statistique centrée réduite), où le signe *prime* (') représente une valeur centrée réduite.

Le test partiel le plus couramment employé dans notre laboratoire est le second, qui a la même valeur qu'une corrélation partielle paramétrique entre les valeurs qui se trouvent dans les matrices **A** et **B**, conditionnellement aux valeurs de la matrice **C**. Seule l'option (2) paraît actuellement acceptable en analyse spatiale (Oden & Sokal, soumis).

(3) Option 4: Le corrélogramme de Mantel

Ce programme peut calculer un corrélogramme de Mantel (Sokal, 1986; Oden & Sokal, 1986). Celui-ci présente sur le programme AUTOCOR l'avantage de permettre le calcul d'un corrélogramme à partir de données multidimensionnelles, puisque le corrélogramme lui-même est calculé sur une matrice de similarités ou de distances produite par SIMIL, matrice qui utilise dans la plupart des cas des données multidimensionnelles; voir Legendre & Fortin (1989) pour un exemple. On obtient le corrélogramme de Mantel en demandant l'option de calcul "0" ainsi que l'option "0" pour la matrice **A**; un test de Mantel est alors calculé pour chacune des classes de distances du corrélogramme.



(4) Statistique utilisée

Le programme peut fournir soit la statistique Z de Mantel, qui est simplement la somme des produits croisés des valeurs correspondantes des deux matrices, à l'exclusion de la diagonale:

$Z = \Sigma \Sigma x_{ij} y_{ij}$ pour toutes les paires de valeurs (i, j) des deux matrices,

soit une forme centrée réduite r de cette statistique, tel que proposé par Smouse, Long & Sokal (1986). Pour calculer cette statistique, on centre et on réduit d'abord les valeurs au sein de chacune des matrices de distance (à l'exclusion de la diagonale), avant de calculer la somme des produits croisés; puis on divise par (n - 1) où n est le nombre de paires de distances considérées dans le calcul. Cette statistique est donc équivalente au calcul d'un coefficient de corrélation de Pearson entre les valeurs des deux matrices (diagonale exclue), si bien que les valeurs obtenues sont situées entre -1 et +1. Que la statistique soit centrée réduite ou non, les probabilités associées sont exactement les mêmes.

(5) Probabilités

Les probabilités peuvent être calculées de deux façons: soit par permutations, ou encore par l'entremise d'une transformation du Z ou du r en une autre statistique, appelée t par Mantel (1967), qui est distribuée de façon asymptotiquement normale centrée-réduite. Ce test donne une bonne approximation de la probabilité lorsque le nombre d'objets est suffisamment grand, si un certain nombre d'autres conditions sont également remplies (voir Mielke, 1978). Lorsque le nombre d'objets est grand, le test par permutations devient très long à réaliser. L'usager a alors la possibilité de demander au programme de ne calculer que le test approximatif; il suffit de ne demander aucune permutation. Pour les tests par permutations, une limite de 200 objets est inscrite dans le programme. **[versions CMS et VMS seulement ?]** Si le nombre d'objets du problème excède cette valeur, le programme ne réalise, d'autorité, que le test par approximation dans le cas de l'option 0 (Mantel entre deux matrices ou corrélogramme simple); le programme s'arrête dans le cas des tests partiels (options 1 à 3). Cette borne est établie par un paramètre en en-tête du programme (PETITNOBJ = 200), que l'usager pourra modifier selon ses besoins dans les versions pour grands ordinateurs.

On a souvent présenté les probabilités des tests de Mantel en tant que surface sous la courbe *à gauche* de la valeur observée; ainsi, une statistique de Mantel négative et significative avait une probabilité près de 0, alors qu'une statistique positive et significative avait une probabilité près de 1. Notre programme présente plutôt la probabilité estimée (test unilatéral) que l'hypothèse nulle (H_0: pas de relation linéaire entre les deux matrices) soit vraie, comme c'est la coutume dans les tests statistiques. Ainsi, les statistiques de Mantel significatives ont une probabilité près de zéro, que la statistique elle-même ait un signe positif ou négatif.

(6) Les tests par permutations

Si l'usager demande d'effectuer le test par permutations, il doit indiquer le nombre de permutations désirées. Si on ne désire pas que le test par permutations soit réalisé, il suffit d'indiquer au programme de réaliser zéro permutations. Lorsque les matrices sont très grandes, le test par permutations devient inutile puisque le test par approximation converge asymptotiquement vers la loi normale centrée-réduite.

Les probabilités obtenues par permutations sont calculées selon la méthode de Hope (1968), méthode recommandée également par Edgington (1987); celle-ci consiste à inclure la valeur observée parmi les valeurs de la distribution de référence, de sorte qu'il n'est jamais possible d'obtenir 0% de valeurs "aussi extrêmes ou plus extrêmes que la valeur observée". Selon Edgington, cette façon de faire est biaisée mais elle a le mérite d'être valide. À tout événement, les probabilités doivent être interprétées en termes de "strictement plus petit" ou "strictement plus grand" que la valeur seuil; ainsi, si la probabilité obtenue par permutation est de 0.05 alors la probabilité que l'hypothèse nulle soit vraie est strictement plus petite que 0.05 dans un test unilatéral. La précision de cette probabilité est toujours

l'inverse du nombre de permutations demandées par l'usager.

(7) Standardisation de Hubert

La standardisation proposée par Hubert (1985), qui produit aussi des valeurs entre -1 et +1, consiste à centrer la valeur réelle de *Z* ou de *r* par rapport aux valeurs extrêmes (minimum et maximum) obtenues au cours des permutations, puis à attribuer à cette statistique le signe qu'avait *Z* ou *r*. Une valeur standardisée selon Hubert égale à +1 signifie essentiellement que la valeur observée de la statistique est la plus grande de la distribution de référence, alors qu'une valeur égale à -1 signifie que la valeur observée est la plus petite de la distribution de référence.

(8) Permutations limitées

Ce programme permet d'effectuer des permutations limitées à des échanges entre les objets membres de certains sous-groupes définis par l'usager (Sokal *et al.*, 1987). L'usager doit indiquer combien de sous-groupes il désire reconnaître, puis il doit donner la liste des objets membres de chaque sous-groupe. Les numéros d'objets peuvent être donnés un par un, ou encore par blocs à l'aide d'un tiret (versions CMS ou VMS seulement); par exemple: 1 4 7 9-32 38 67 serait une réponse valide. Si on ne désire pas se prévaloir de cette option, on répond "1" à la question "Nombre de groupes à permuter ? (Problème général: 1)". Le principe de ce test est expliqué ci-dessous.

Considérons le cas d'un test de conformité d'un modèle à des données. La méthode consiste à formuler l'hypothèse alternative (H_1) — par exemple, l'existence de groupes distinguables dans les données — sous la forme d'une matrice-modèle, contenant par exemple des "1" entre les objets supposés appartenir au même groupe et des "0" ailleurs. Une matrice de ressemblance est également calculée pour les données. L'hypothèse nulle (H_0) de non-conformité du modèle aux données est testée en comparant la valeur de la statistique de Mantel à une distribution de référence obtenue par permutations successives de l'une des matrices suivies du re-calcul de la statistique de Mantel.

Si deux hypothèses alternatives concurrentes d'appartenance à des groupes sont toutes deux significativement conformes aux données (matrice **A**), on peut procéder comme suit pour évaluer si l'une des deux rend mieux compte des données:

1- Exprimer chacune des deux hypothèses alternatives sous la forme d'une matrice-modèle, que nous appellerons $\mathbf{B_1}$ et $\mathbf{B_2}$. Les paires d'objets groupés par l'hypothèse alternative 1 reçoivent des "1" dans la matrice $\mathbf{B_1}$ alors que celles qui sont groupées par l'hypothèse 2 reçoivent des "1" dans $\mathbf{B_2}$.

2- On réalise un test de Mantel entre la matrice **A** correspondant aux données et la matrice-modèle $\mathbf{B_1}$, en ne permutant qu'à l'intérieur des groupes reconnus par la seconde hypothèse alternative; $\mathbf{B_2}$ devient en fait l'hypothèse nulle de ce test.

3- De même, on réalise un test de Mantel entre la matrice **A** correspondant aux données et la matrice-modèle $\mathbf{B_2}$, en ne permutant qu'à l'intérieur des groupes reconnus par l'hypothèse alternative 1; $\mathbf{B_1}$ devient en fait l'hypothèse nulle de ce test.

4- Si un seul test demeure significatif, on retient l'hypothèse alternative qui lui correspond.

L'exemple suivant a été traité par Legendre & Lessard (en prép.). La question est de savoir si des filets de maille différente pêchent essentiellement les mêmes espèces de poissons à une série de stations d'échantillonnage. L'hypothèse nulle est que les différences entre échantillons sont indépendantes des stations ou des types de filets. La première hypothèse alternative est que les deux types de filets échantillonnent la même communauté de poissons à chaque station; si cette hypothèse est supportée par les données, il devient possible de regrouper les résultats de pêche par les deux types de filets, pour l'études des communautés de poissons. La seconde hypothèse alternative affirme au contraire que le premier filet échantillonne une première communauté (petites espèces) à toutes les

stations, alors que le second filet, de plus grande maille, échantillonne une seconde communauté (espèces plus grandes). Ces trois hypothèses peuvent être représentées par les vecteurs suivants; les chiffres représentent des types de communautés, en supposant qu'il y a 5 stations d'échantillonnage:

	Station 1		Station 2		Station 3		Station 4		Station 5	
	Filet 1	Filet 2	Filet 1	Filet 2	Filet 1	Filet 2	Filet 1	Filet 2	Filet 1	Filet 2
Observation no	1	2	3	4	5	6	7	8	9	10
Hypothèse nulle	1	1	1	1	1	1	1	1	1	1
Hypothèse alternative 1	1	1	2	2	3	3	4	4	5	5
Hypothèse alternative 2	1	2	1	2	1	2	1	2	1	2

Chacun de ces vecteurs de nombres peut aisément être transformé en une matrice-modèle par le calcul d'un coefficient de Jaccard pour données multiclasses, à l'aide des coefficients S15 ou S16 du programme SIMIL. Dans le premier test par permutations limitées, on teste entre la matrice **A** (similarités basées sur les données réelles) et la matrice-modèle $\mathbf{B_1}$ en ne permutant qu'à l'intérieur des groupes-filets (1-3-5-7-9) et (2-4-6-8-10). Ensuite, on réalise le test de Mantel entre **A** et la matrice-modèle $\mathbf{B_2}$ en ne permutant qu'à l'intérieur des groupes-stations (1-2), (3-4), (5-6), (7-8) et (9-10).

Les questions du programme

La première question du programme concerne le type de calcul. On doit choisir entre le test de Mantel simple ou l'un des tests partiels. Si on désire un corrélogramme de Mantel, celui-ci est le plus souvent réalisé par l'option 0 (Mantel simple); le programme permettrait cependant de calculer un corrélogramme partiel, c'est-à-dire un corrélogramme fait de tests de Mantel partiels. Voir la façon de spécifier les classes de distances dans le fichier d'entrée **A**, au point (2.2) ci-dessus.

La seconde question concerne les options pour la matrice **A**. Ces options sont décrites en (2) ci-dessus. Si on indique au programme que les points forment une grille régulière, la question suivante concerne la largeur de cette grille (nombre de colonnes); le nombre total de points étant connu du programme (puisqu'il est inscrit dans l'en-tête binaire du fichier **B** produit par *SIMIL*), le nombre de lignes de la grille est calculé automatiquement. Si par ailleurs on a choisi de fournir au programme un fichier de coordonnées géographiques (options 3 ou 4), une question subséquente du programme permettra de spécifier comment seront calculées les distances: soit par la formule de la distance euclidienne (coordonnées planes), soit en suivant la courbure de la terre (coordonnées sur une sphère); dans ce dernier cas, les distances sont exprimées en milles marins.

On devra ensuite déterminer si on désire obtenir la statistique Z de Mantel originale, ou encore la statistique centrée réduite r qui est bornée entre les valeurs -1 et +1. Cette question n'est cependant pas posée lorsqu'on a demandé un test partiel de Smouse, Long & Sokal, car la statistique dans ce cas doit être une corrélation partielle.

L'usager doit maintenant préciser combien de permutations il désire obtenir; s'il en demande "zéro", seul le test par approximation est réalisé. Les utilisateurs du test de Mantel demandent souvent 999 permutations (pour un total de 1000 avec la valeur réelle); il est cependant recommandé d'accroître substantiellement ce nombre lorsqu'on s'approche du seuil de signification α pré-établi, à cause de l'instabilité des probabilités obtenues par la méthode des permutations (Jackson & Somers, 1988).

La dernière question concerne le nombre de groupes à permuter; voir la section (8) ci-dessus. Si ce nombre est différent de 1, le programme demandera le nombre d'objets appartenant à chaque groupe, puis les numéros de ces objets; on doit fournir les numéros séquentiels des objets dans la

matrice d'entrée, et non les noms attribués aux objets dans les 10 premières colonnes de cette matrice.

Exemple

L'exemple ci-dessous illustre l'utilisation du programme pour calculer une relation partielle de Smouse, Long & Sokal (1986) sur grand ordinateur (système CMS ou VMS; cet exemple a été réalisé sous CMS). Le programme de lancement demande d'abord à l'usager d'identifier les fichiers qui seront utilisés; les réponses sont soulignées. Puis, après l'en-tête, viennent les questions posées par le programme lui-même pour identifier quelles sont les options de calcul que désire l'usager. Cet exemple est l'un des résultats rapportés par Legendre & Troussellier (1989): il s'agit du test de Mantel partiel entre les variables MA et CHLA sur l'étang de Thau, en contrôlant l'effet de la matrice des distances géographiques (XY).

```
Programme MANTEL3, decembre 1989.

Ce programme utilise 2 ou 3 matrices de distances; deux pour les
tests de Mantel (elles s'appellent alors A et B), et trois pour
les tests de Mantel partiels (elles s'appellent alors A, B et C).
On peut employer des matrices de similarites plutot que des
matrices de distances, mais il n'est pas recommande de meler les
types, ce qui compliquerait inutilement l'interpretation
MATRICE "A" :
La matrice "A" peut etre une matrice de distances de type
SIMIL. Si tel est le cas, quel est le nom du fichier qui la
contient? (Par defaut: "... data a")
MA  D01  B

COORDONNEES DES POINTS :
Si les points forment une grille rectangulaire reguliere, il
n'est besoin d'aucun fichier pour calculer la matrice "A" des
distances geographiques entre ces points. Dans le cas contraire,
et si vous n'avez pas declare ci-dessus de fichier pour la
matrice "A", le programme aura besoin d'un fichier contenant
les coordonnees des objets. Quel est le nom de ce fichier, s'il y
a lieu? (Par defaut: "... data a")

MATRICE "B" :
"B" est une matrice de distances de type SIMIL. Quel est le nom du
fichier qui la contient? (Par defaut: "... data a")
CHLA  D01  B

MATRICE "C" :
Si le calcul demande une matrice "C", celle-ci est aussi une
matrice de distances de type SIMIL. Quel est le nom du fichier
qui la contient, s'il y a lieu? (Par defaut: "... data a")
XY  D01  B

CLASSES DE DISTANCE :
Pour calculer un correlogramme de Mantel, le programme aura besoin
d'un fichier contenant une matrice triangulaire superieure de
classes de distances; cette matrice se presente sans la diagonale.
Quel est le nom de ce fichier, s'il y a lieu?
(Par defaut: "... data a")
P R O G R A M M E   M A N T E L   avec test par permutations
```

```
Auteur: A. Vaudor
Departement de Sciences biologiques, Universite de Montreal,
C.P. 6128, Succursale a, Montreal, Quebec H3C 3J7.

Type de calcul:
   (0) Mantel entre deux matrices
   (1) Dow & Cheverud (A.(B-C))
   (2) Smouse, Long & Sokal (AB.C)
   (3) Hubert (A.(BC))
2

Options pour la matrice "A":
   (0) Fichier d'entree en classes (pour correlogramme)
   (1) Grille reguliere (aucun fichier n'est requis)
   (2) Fichier de distances (ou de similarites) de simil
   (3) Fichier de coordonnees en degres, minutes et secondes
   (4) Fichier de coordonnees en degres decimaux
2

Nombre d'iterations ? -- (Recommande >= 250)
999

Nombre de groupes a permuter ? (Probleme general: 1)
1

Test unilateral a gauche ou a droite:
Les probabilites sont significatives pres de zero.
PP signifie Plus Petits, EG EGaux et PG, Plus Grands que la stat. originale.
La valeur originale est ajoutee aux EGaux, suivant Hope (1968).


Calcul:                                        Permutations     Approximation
                r     r stand.    PP   EG   PG     Prob(r)       t     Prob(t)
                     **Hubert**                  (Hope,1968)

  AB.C       0.25210   0.96420  997    1    2     0.00300   4.19588   0.00001

Fin du programme.
```

Contenu du fichier de résultats (version Macintosh)

La version Macintosh inscrit les résultats dans un fichier, alors que les versions CMS et VMS les présentent plutôt à l'écran, comme on l'a vu dans l'exemple ci-dessus. Le fichier rappelle d'abord quelles sont les matrices de ressemblance qui ont été utilisées pour le calcul, en reproduisant le bloc d'informations inscrites (en binaire) au début de chacun des fichiers produits par *SIMIL*. Dans le cas d'un test de Mantel simple, il n'y a pas d'identification de méthode à la gauche de la ligne des résultats.

L'exemple qui suit a été calculé sur Macintosh. Les résultats ci-dessus nous apprennent que la relation de Mantel (r = 0.25210) est positive et significative au seuil α = 5% ($p_{1000\ \text{permutations}}$ = 0.003, $p_{\text{approximation}}$ = 0.00001). Il s'agit du test de Mantel simple entre les matrices MA et XY, traitées également ci-dessus. Le détail des permutations est rapporté: 997 permutations ont conduit à des valeurs de la statistique inférieures (PP) à la valeur obtenue pour les matrices originales; aucune

des valeurs de la statistique inférieures (PP) à la valeur obtenue pour les matrices originales; aucune valeur obtenue par permutation n'était égale à la vraie valeur, puisque le nombre rapporté sous "EG" comprend d'abord la valeur elle-même, suivant la méthode de Hope. Enfin, 2 résultats obtenus par permutations étaient supérieurs à la vraie valeur. La probabilité estimée par les permutations est obtenue par (EG + PG)/(nombre de permutations + 1) = 3/1000 dans cet exemple. Pour un test unilatéral à gauche, cette probabilité serait calculée par (PP + EG)/(nombre de permutations + 1). Notez qu'avec un problème de cette taille (63 observations), il n'aurait pas été nécessaire de procéder par permutations, les résultats obtenus par le test *t* approximatif se rapprochant suffisamment des résultats permutationnels si le seuil de signification pré-établi est de 5%. Si on s'intéressait plutôt à un seuil de 0.001, il faudrait alors augmenter substantiellement le nombre de permutations, d'abord pour minimiser l'effet de la correction de Hope, ensuite afin de vérifier de quel côté du seuil tombe le résultat.

```
Nombre d' itérations:  999
Test unilatéral à gauche ou à droite:
Les probabilités sont significatives près de zéro.
PP signifie plus petits, EG égaux et PG, plus grands que la stat. originale.
La valeur originale est ajoutée aux EGaux, suivant Hope (1968).

Calcul:                                        Permutations     Approximation

               r      r stand.  PP   EG   PG     Prob(r)       t     Prob(t)
                    --Hubert--                (Hope,1968)

            0.22338   1.00000  999    1    0     0.00100   4.69498   0.00000
```

Dans un corrélogramme, une ligne de résultats est présentée pour chacune des classes de distances demandées. L'exemple suivant a été calculé sur le Macintosh.

```
**** Notez que dans ce corrélogramme, l'autocorrélation positive
     produit des z négatifs à faible distance

Nombre d' itérations:  249
Test unilatéral à gauche ou à droite:
Les probabilités sont significatives près de zéro.
PP signifie plus petits, EG égaux et PG, plus grands que la stat. originale.
La valeur originale est ajoutée aux EGaux, suivant Hope (1968).

Calcul:                                        Permutations     Approximation

               r      r stand.  PP   EG   PG     Prob(r)       t     Prob(t)
                    --Hubert--                (Hope,1968)

classe   1
           -0.19512  -0.50163   30    1  219     0.12400  -1.26150   0.10356
classe   2
           -0.23068  -0.58478   13    1  236     0.05600  -1.60066   0.05473
classe   3
           -0.22218  -0.64604   17    1  232     0.07200  -1.60212   0.05457
classe   4
            0.07324   0.17663  178    1   71     0.28800   0.48162   0.31504
classe   5
            0.12409   0.28565  203    1   46     0.18800   0.87896   0.18971

classe   6
```

```
             0.12082   0.32244  191    1   58     0.23600   0.73562   0.23098
classe   7
             0.24780   0.59178  230    1   19     0.08000   1.50877   0.06568
classe   8
             0.38124   1.00000  245    5    0     0.02000   2.25543   0.01205
```

Notez que dans ce type de corrélogramme, si la matrice **B** est une matrice de distances, la statistique de Mantel aura un signe négatif en cas d'autocorrélation positive; c'est l'inverse si la matrice **B** est plutôt de type similarités, le signe positif indiquant alors la présence d'autocorrélation positive. Tel est le sens de la note initiale. Pour tracer le corrélogramme, il suffira de porter les valeurs de *r* en ordonnée en fonction des classes de distances en abscisse. La signification pourra être prise soit dans la colonne du test par permutations, soit dans celle du test approximatif; il convient d'utiliser la correction de Bonferroni pour évaluer le degré de signification d'un tel corrélogramme, tel que recommandé également pour les programmes AUTOCORRÉLATION SPATIALE et PÉRIODOGRAPHE. Dans cet exemple où la matrice **B** était une matrice de distances, il convient de changer tous les signes des statistiques de Mantel avant de tracer le corrélogramme.

PCOORD

Que fait PCOORD ?

Ce programme produit une ordination en espace réduit par la méthode des coordonnées principales (Gower, 1966). Comme l'analyse en composantes principales, cette méthode réalise un *cadrage multidimensionnel métrique.* Cependant, les calculs sont effectués à partir d'une matrice de *similarités* ou de *distances* plutôt qu'à partir d'un tableau de données brutes; tel est également le cas avec les méthodes de *cadrage multidimensionnel non-métrique* (*"nonmetric multidimensional scaling"* en anglais).

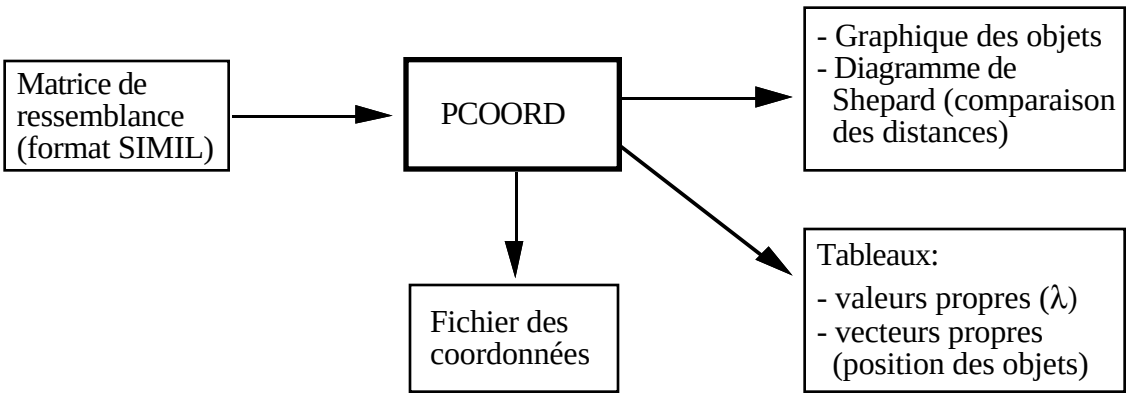
Chaque distance *d* est d'abord transformée en une nouvelle distance $d' = -d^2/2$ avant d'effectuer un centrage par la formule

$$alpha = d' - d'bar_i - d'bar_j + d'bar$$

où $d'bar_i$ et $d'bar_j$ sont respectivement la moyenne de la ligne *i* et de la colonne *j* dans la matrice de distances *d'*, alors que *d'bar* est la moyenne de toutes les valeurs de la matrice. Les nouvelles coordonnées des objets dans l'espace réduit sont les vecteurs propres de cette matrice centrée, après normalisation à la racine carrée de leur valeur propre.

La taille des matrices de distances qui peuvent être traitées par ce programme est limitée, dans les versions CMS et VMS, par le paramètre DIMENSION au début du programme. Si cette limite s'avère insuffisante, il suffit de recompiler le programme après avoir modifié cette constante. Il n'y a en principe pas de limite quant à la taille des matrices de ressemblance qui peuvent être traitées par la version Macintosh du programme. Le programme occupe tout l'espace mémoire (RAM) qui lui est disponible, si bien que la taille des matrices que le programme peut traiter en pratique sera une fonction, non seulement de la taille de la mémoire disponible dans la machine, mais également de l'utilisation simultanée de MultiFinder, d'une mémoire-cache ou d'autres programmes. Sous le système 6.04, des matrices de taille **???** peuvent être traitées avec 1 Meg de mémoire RAM.

Fichiers d'entrée et de sortie



(1) Fichier d'entrée

Le fichier d'entrée est une matrice (p x p) de similarités ou de distances de type binaire tel que produite par les programmes SIMIL, IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh) décrivant la ressemblance entre p objets pour n variables.

En principe, la matrice soumise à ce calcul doit correspondre à une distance *métrique* permettant

une représentation *euclidienne* des objets. Dans ce cas, p objets produiront au maximum (p - 1) valeurs propres positives et une valeur propre nulle, puisqu'il suffit de (p - 1) dimensions pour représenter la position relative de p objets dans un espace euclidien. Des valeurs propres négatives sont produites lorsque les distances entre objets ne peuvent être entièrement représentées de façon euclidienne. Gower (1982) a montré que dans certains cas, une mesure de distances métriques peut produire une représentation non-euclidienne des objets, alors que Gower et Legendre (1986) ont décrit les conditions permettant une représentation euclidienne d'une matrice de ressemblance. Quoiqu'il en soit, la fraction non-euclidienne de la représentation en espace réduit n'a pas beaucoup d'importance pourvu qu'elle soit, en valeur absolue, nettement moins importante que la variabilité exprimée par les premières coordonnées principales.

(2) Fichiers des résultats

Ce fichier contient les valeurs propres et la position des objets par rapport aux trois premiers vecteurs propres, si ceux-ci ont une valeur propre positive. Si certaines des valeurs propres sont négatives, le pourcentage de variance correspondant à chaque valeur propre λ_i est corrigé par la valeur absolue de la plus grande valeur propre négative (Cailliez & Pagès, 1976), par la formule

$$\lambda_i' = (\lambda_i + |\lambda_p|) / [\sum_{j=1}^{p-1} \lambda_j + (p - 1) |\lambda_p|]$$

où $|\lambda_p|$ est la valeur absolue de la plus grande valeur propre négative et p est le nombre d'objets; p représente également le nombre de valeurs propres calculées.

Trois graphiques des objets sont ensuite fournis: un premier pour les axes I et II, un second pour I et III et un troisième pour II et III. Dans les versions CMS et VMS, ces graphiques, qui sont imprimés latéralement, présentent le premier axe verticalement et le second pointant vers la gauche. Cette représentation est justifiée par le fait que puisque l'axe I est toujours plus variable que l'axe II, il sera donc probablement plus long. Tournez ces graphiques de 90° avant reproduction.

(3) Fichier des coordonnées

Un second fichier, nommé par défaut "COORD data a" dans la version CMS et écrit en ASCII, contiendra la position des objets par rapport à autant d'axes principaux que requis par l'usager; ce nombre ne peut cependant être plus grand que le nombre de valeurs propres positives. Si on désire obtenir des graphiques de plus de trois axes principaux, il est facile de transférer ce fichier à un micro-ordinateur où il pourra être traité par un logiciel statistique. Ce fichier pourra également servir de base à un groupement non-hiérarchique par la méthode *k-means*, tel qu'expliqué au chapitre portant sur le programme *K-MEANS*.

Ces deux mêmes types de fichiers (2 et 3) peuvent être produits par la version Macintosh. Les graphiques sont de qualité de publication, comme on peut le voir dans les exemples ci-dessous; l'usager peut obtenir toutes les combinaisons d'axes qu'il désire. Les graphiques sont d'abord présentés à l'écran; on peut les faire imprimer, ou encore les conserver dans un fichier de type PICT pour usage futur. La version Macintosh peut également produire un diagramme de Shepard (diagramme de dispersion comparant les distances d'origine et les distances dans l'espace de dimension réduite: voir l'exemple).

Les questions du programme

Les questions présentées par le programme à l'écran du Macintosh sont décrites dans les paragraphes qui suivent. Les questions posées par les versions CMS et VMS sont essentiellement les mêmes, comme on le verra dans l'exemple ci-dessous, sauf pour ce qui est des diagrammes de

mêmes, comme on le verra dans l'exemple ci-dessous, sauf pour ce qui est des diagrammes de Shepard qui ne sont pas disponibles dans les versions du programme pour grands ordinateurs. Pour faire démarrer le programme sur le Macintosh, il faut cliquer sur l'icône, puis donner la commande "Ouvrir" dans le menu "Fichier".

(1) "Titre de ce travail ..." — On fournit un titre, qui sera repris en en-tête des graphiques reproduits par l'imprimante.

(2) "Est-ce une matrice de distances plutôt que de similarités?" [Oui, Non] — On répond *Oui* s'il s'agit d'une matrice de distances.

(3) "Fichier d'entrée" — Le programme présente le menu des fichiers de type *SIMIL* disponibles, puisque la matrice de ressemblance soumise à ce programme doit provenir soit du programme SIMIL, soit du programme IMPORT-EXPORT (version Macintosh) ou IMPORT (versions CMS et VMS).

(4) "Combien de valeurs propres à extraire?" — L'algorithme utilisé dans la version Macintosh pour le calcul des valeurs propres est un algorithme pas à pas **[nom?]** qui calcule d'abord les valeurs propres les plus importantes. L'usager peut limiter le calcul aux quelques premières valeurs propres (habituellement de 2 à 5) qui contiennent habituellement la plus grande partie de la variance; cela peut représenter un gain de temps appréciable pour les problèmes comportant de nombreux objets.

(5) "Combien de dimensions voulez-vous dessiner?" — Des graphiques successifs seront produits pour toutes les paires d'axes principaux demandés. Ainsi, si on demande de représenter 3 dimensions, trois graphiques seront produits, correspondant respectivement aux axes I et II, I et III, II et III. Pour augmenter la résolution, on peut agrandir n'importe quelle partie de l'image en l'entourant d'un cadre à l'aide de la souris.

(6) "Numérotation sur le graphique?" [Oui, Non] — Si on répond *Oui*, des numéros séquentiels permettent d'identifier les objets sur chacun des graphiques.

La liste des valeurs propres est disponible dans le menu "Calculs (détails)". On peut monter ou descendre dans ce tableau en pointant le curseur de la souris dans le bas ou le haut du tableau. Ces résultats peuvent être envoyés directement à l'imprimante ou copiés dans un fichier de résultats pour référence future. De même, à partir du menu "Graphiques", les graphiques peuvent être envoyés à l'imprimante, ou encore on peut les préserver dans un fichier de type PICT, ce qui permettra de les éditer à l'aide d'un programme graphique ou de les inclure dans un fichier de texte. Il faut "Terminer" chaque graphique pour passer au suivant, ou encore pour passer à la question suivante.

(7) "Dans combien de coordonnées désirez-vous réécrire les objets?" — L'usager indique combien de coordonnées (nombre entier) il désire voir écrire sur un fichier; un nom de fichier lui sera demandé. On répond "0" (zéro) si on ne désire pas ce fichier.

(8) "Comparaison des distances?" [Oui, Non] — Cette question n'apparaît pas dans les versions CMS et VMS. Si on répond *Oui*, les questions suivantes permettent de préciser comment se fera la comparaison (diagramme de Shepard) entre les distances dans la matrice de ressemblance d'origine et les distances dans l'espace réduit à 2, 3, … dimensions. Dans ce graphique, un nuage de points étroit, situé sous la diagonale mais près de celle-ci, indique une bonne représentation des distances d'origine dans l'espace réduit. Si on a utilisé une distance qui ne peut être entièrement représentée dans un espace euclidien, des points pourront apparaître au-dessus de la diagonale.

(8.1) "Nombre de valeurs propres à comparer?" — On indique combien de dimensions de l'espace réduit seront incluses dans cette comparaison des distances (en général, 2 ou 3).

(8.2) "XX distances à calculer; préférez-vous échantillonner celles-ci?" [Oui, Non] — Il y a XX = $p(p\text{-}1)/2$ distances entre p objets. Lorsque ce nombre devient trop grand (plus de quelques centaines:

calcul trop long), l'usager peut demander à l'ordinateur de choisir au hasard un certain nombre de ces distances. Le nombre en question sera déterminé à la question (8.3), la sélection étant réalisée au hasard à l'aide d'un générateur de nombres pseudo-aléatoires initialisé à la question (8.4).

(8.3) "Combien de distances à échantillonner?" — On inscrit le nombre de distances désiré.

(8.4) "Générateur de nombres aléatoires: entrer un (petit) chiffre" — On inscrit un petit entier positif, par exemple 2, 5 ou 10.

(8.5) "Autre comparaison des distances?" [Oui, Non] — Si on répond *Oui* à cette question, on retourne à la question (8.1). La réponse *Non* met fin à l'exécution du programme.

Exemple

Dans l'exemple qui suit, un fichier de distances de Mahalanobis a été calculé au préalable entre 9 groupes d'observations. Le fichier de distances calculé par SIMIL porte le nom de **mahal d5 a**; il sert de fichier d'entrée lors de cette analyse en coordonnées principales. L'exemple ci-dessous a été calculé sous CMS. Le fichier d'appel CMS ou VMS pose les questions suivantes, qui sont suivies des questions posées par le programme.

Noter la question portant sur la largeur des graphiques (point 1, en marge gauche): si par exemple on désire un graphique de 8 pouces de largeur (20 cm), on répond "8" à cette question; l'algorithme utilisé exige que toutes les réponses soient des multiples de 4. La matrice utilisée étant une matrice de distances, on répond **d** à la question (2). Enfin (3), on demande que la position des objets par rapport aux **5** premières coordonnées principales soit écrite sur fichier.

```
     Pcoord
     Quel est le nom du fichier de type SIMIL?
     (Par defaut: "... data a")
     mahal  d5  a

     Quel nom doit recevoir le fichier de sortie (valeurs propres
     et graphiques)?  (Par defaut: "... listing a")
     mahal  sortie  a

     Quel nom doit recevoir le fichier de sortie COORD (contenant les
     coordonnees des objets), s'il y a lieu?
     (Par defaut: "COORD data a")

     Execution begins...

(1)   LARGEUR DU GRAPHIQUE? (en pouces: multiples de 4)
     8
      LA MATRICE D'ENTREE EST DE SIMILARITES OU DE DISTANCES? (S ou D)
(2)  d
      REECRITURE DES OBJETS DANS LE NOUVEL ESPACE (Fichier "COORD"):
      COMBIEN DE COORDONNEES VOULEZ-VOUS ?  (0 si aucun)
(3)  5
      TITRE DU TRAVAIL ?
     Distances  de  Mahalanobis,  9  groupes

      Fin du programme.
```

Graphiques et contenu du fichier de résultats

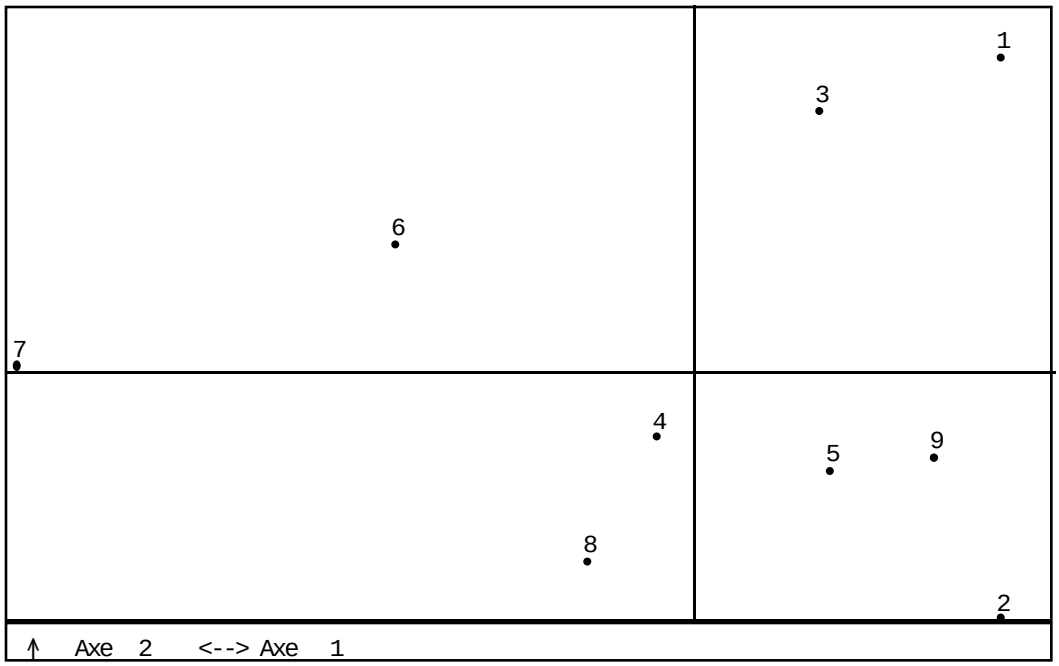
Le premier fichier contient les valeurs propres, ainsi que le pourcentage de variance expliqué par chacune. Puisqu'il y a des valeurs propres négatives, la correction décrite à la section du fichier de résultats a été utilisée.

```
Valeurs propres       % de variance
        9.43558            50.40291
        3.55587            18.99487
        3.06849            16.39137
        1.62149             8.66186
        0.70898             3.78740
        0.30302             1.61886
        0.02664             0.14252
       -0.00000             0.00022
       -0.00004             0.00000
```

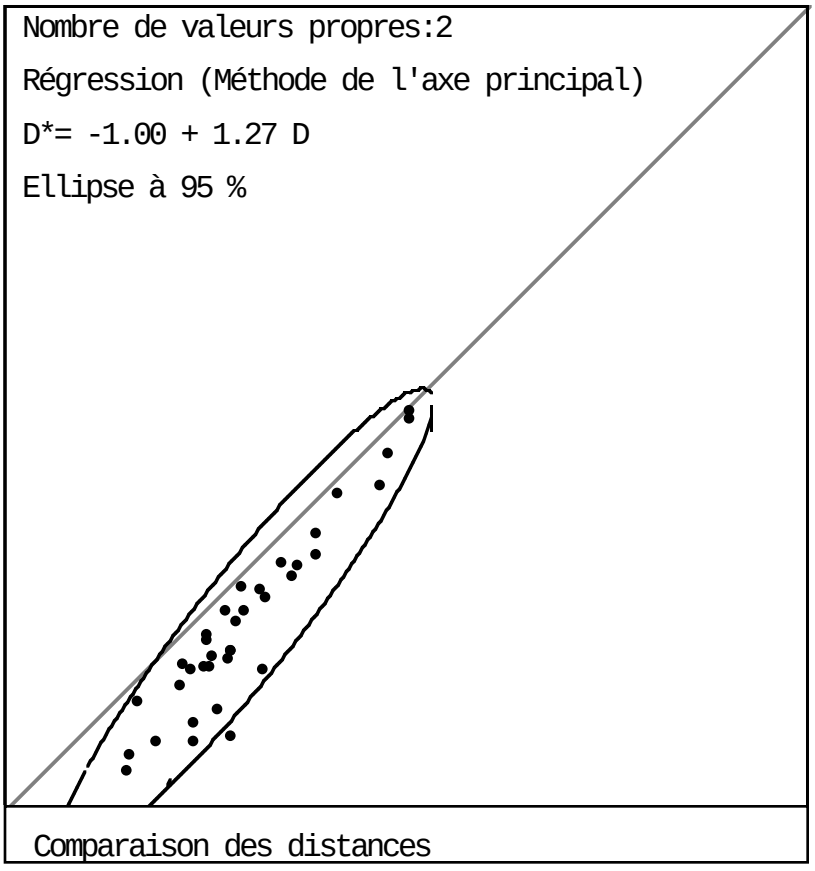
Un autre fichier contient le nombre de coordonnées principales que l'on a demandé d'écrire (ici, 5 coordonnées). Chaque ligne de ce fichier représente donc les coordonnées d'un objet par rapport à 5 dimensions.

```
 1.03108   1.06888   0.80571   0.12374  -0.27798
 1.03064  -0.85473   0.17086   0.84875   0.15406
 0.42006   0.88857  -0.96345   0.28882   0.08743
-0.12718  -0.23230  -0.08304  -0.38978  -0.45709
 0.45717  -0.34516   0.69916  -0.44539   0.31628
-0.99884   0.43152  -0.34929  -0.14860   0.41268
-2.26578   0.01133   0.59580   0.27563  -0.03772
-0.35435  -0.66460  -0.66777   0.04047  -0.32255
 0.80720  -0.30352  -0.20799  -0.59363   0.12490
```

Le graphique de la position des objets dans les deux premières dimensions est présenté ci-après (version Macintosh):



Le diagramme de Shepard ci-dessous, qui compare les distances d'origine (abscisse) aux distances dans l'espace des deux premières coordonnées principales (ordonnée), fait état d'un nuage de points étroit et près de la diagonale; cela nous indique que les distances d'origine sont bien représentées par deux dimensions seulement.



*PÉRIODOGRAPHE*Macintosh *ou PERIOD*CMS/VMS

Que fait le PÉRIODOGRAPHE ?

Ce programme calcule et trace un périodogramme de contingence (Legendre *et al.*, 1981) pour une série temporelle ou spatiale de données unidimensionnelles. Les données peuvent être qualitatives (nominales), semi-quantitatives (ordinales) ou quantitatives. Les données quantitatives et semi-quantitatives doivent d'abord être divisées en classes avant le calcul de ce périodogramme; le programme se charge de réaliser cette division selon un critère d'optimisation. Pour le périodogramme, le programme calcule la statistique de contingence pour toutes les périodes comprises dans la fenêtre d'observation, soit les périodes de $T = 2$ à $T = n/2$ où n est la longueur de la série; dans la version Macintosh du programme, l'usager peut choisir une fenêtre de calcul plus étroite. Legendre & Legendre (1984a, tome 2, pages 228-231), de même que l'article cité ci-dessus, fournissent plus de détails sur la méthode. Outre sa capacité d'analyser des séries de données semi-quantitatives ou qualitatives, la méthode présente également l'avantage de permettre l'analyse de séries courtes, ce qui n'est pas le cas avec le périodogramme de Schuster ou l'analyse spectrale, par exemple. Pour l'analyse de séries multidimensionnelles, on préférera calculer un corrélogramme de Mantel (voir ce programme) plutôt qu'un périodogramme de contingence après classification multivariable des données, tel que nous l'avions proposé dans l'article de 1981; de plus, la méthode du corrélogramme de Mantel ne requiert pas que le pas d'échantillonnage soit régulier.

La division d'une variable quantitative ou semi-quantitative en classes est réalisée à l'aide d'une procédure qui optimise les deux critères suivants, de façon à tenir compte des valeurs liées (*ex aequo*) dans la série des données:

1- pour un nombre de classes donné, on minimise la somme des variabilités intra-classes (calcul effectué sur les valeurs brutes ou sur les rangs);
2- on cherche le nombre de classes qui maximise la quantité d'entropie par classe formée.

Un algorithme pas à pas, transcrit dans la procédure APPROX du programme, est décrit dans l'article de Legendre *et al.* (1981: 969-973); dans cette procédure, on cherche d'abord la division en deux classes qui minimise le premier critère puis, gardant cette première division fixe, on cherche un second point de coupure qui crée trois classes minimisant de nouveau le critère de variance, et ainsi de suite jusqu'à maximisation du second critère. Un second algorithme a été récemment mis au point par A. Vaudor; cette méthode, traduite dans la procédure EXACT du programme, trouve à chaque étape la partition optimale des observations en k classes, et ce indépendamment des bornes de classes trouvées à l'étape précédente; la partition qui maximise l'information par classe est retenue. Le programme emploie la procédure EXACT toutes les fois où cela est possible. Notons que dans ces algorithmes, le second critère trouve souvent son optimum pour trois classes. L'usager peut toujours imposer au programme de calculer un autre nombre de classes s'il le désire.

Fichiers d'entrée et de sortie

(1) Fichier d'entrée

Le fichier d'entrée est un fichier en caractères lisibles (ASCII) qui peut contenir soit des classes (catégories), soit des séries de valeurs entières ou réelles. La version Macintosh impose les limites suivantes: pas plus de 2 000 valeurs réelles, ou 10 000 valeurs entières, ou encore 60 classes. Dans

les versions CMS et VMS, l'usager fixe lui-même les paramètres du programme qui déterminent ces limites, avant la compilation; le paramètre LIMITE établit le nombre maximum de valeurs que l'on peut traiter dans une série de données, alors que le paramètre LIMCLASSES fixe le nombre maximum de classes dans chaque série de données qualitatives.

Les observations sont entrées dans leur ordre temporel ou dans le sens du transect pour des données spatiales, sans identificateur, une série après l'autre. Il y a trois points à vérifier à propos du fichier d'entrée:

1- Toutes les données doivent être strictement positives. Cette restriction vient du fait que la méthode a d'abord été mise au point pour des données nominales, codées en *k* classes numérotées habituellement de 1 à *k*. Si on désire analyser des données quantitatives comportant des valeurs nulles ou négatives, il faut les transformer avant de les soumettre à ce programme; il est facile de rendre des données strictement positives à l'aide du programme VERNORM, ou encore à l'aide des nombreux logiciels statistiques disponibles sur micro-ordinateur.

2- Si on désire analyser simultanément plusieurs séries de données, chaque série doit former une ligne du fichier de données, ou encore être écrite sur une série de lignes consécutives. Toutes les séries analysées dans une seule passe doivent être de même longueur. Il est facile de transposer un fichier de données à l'aide du programme VERNORM, si nécessaire.

3- Comme avec les autres méthodes d'analyse des séries temporelles, le programme suppose que les données sont stationnaires (i.e., même moyenne et même variance pour différentes portions de la série) et que le pas d'échantillonnage (i.e., l'intervalle entre les observations) est constant. Si tel n'est pas le cas, on peut le rendre constant par interpolation. Ce programme ne peut traiter les absences d'informations; celles-ci doivent également être comblées par interpolation ou par une autre forme d'estimation.

Le fichier suivant, qui contient 2 séries de 16 observations, serait un fichier acceptable pour le programme PÉRIODOGRAPHE:

```
1 1 2 3 3 2 1 2 3 2 1 1 2 3 3 1
2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 3
```

(2) Fichier de résultats

Le fichier de sortie contient les informations concernant la division des variables quantitatives ou semi-quantitatives en classes, ainsi que les détails du périodogramme de contingence. Voir l'exemple ci-dessous. Cette sortie apparaît à l'écran seulement dans les versions CMS et VMS; il est possible de la faire transcrire dans un fichier de "mémoire de console" en suivant la procédure de la page 2.

En plus de ce fichier, la version Macintosh produit également des graphiques (périodogrammes) illustrés plus bas. Cette option graphique n'est pas disponible dans les versions CMS et VMS.

Les questions du programme

Les questions posées par les versions CMS et VMS du programme sont illustrées à la section suivante (Exemple). Ces questions sont essentiellement les mêmes que celles qui apparaissent à l'écran du Macintosh, quoique leur formulation puisse légèrement différer dans certains cas. Pour faire démarrer le programme sur le Macintosh, il faut cliquer sur l'icône, puis donner la commande "Ouvrir" dans le menu "Fichiers".

(1) "Fichier de résultats" — Le programme présente un menu permettant de nommer le fichier appelé à contenir les résultats des calculs. Un nom est suggéré par défaut.

contenir les résultats des calculs. Un nom est suggéré par défaut.

(2) “Fichier de données” — Le programme présente le menu des fichiers ASCII disponibles.

(3) “Nombre d’observations” — On inscrit le nombre d’observations présentes dans chacune des séries de données.

(4) “Nombre de variables” — On inscrit le nombre de séries de données à analyser.

(5) “Le fichier est-il déjà en classes (Données qualitatives)? [Oui, Non] — On répond *Oui* s’il s’agit de données qualitatives (nominales). Si on répond *Non*, les questions suivantes apparaissent à l’écran:

(5.1) “Nombre de classes? (0 pour calcul par le programme)” — L’usager peut imposer le nombre de classes qu’il désire obtenir, satisfaisant le premier critère de l’algorithme de division en classes (minimisation de la somme des sommes de carrés d’écarts intra-classes), tel qu’expliqué au paragraphe d’introduction ci-dessus. S’il répond “0”, le programme déterminera le nombre optimum de classes selon le deuxième critère de l’algorithme (maximisation de la quantité d’entropie par classe).

(5.2) “Calcul sur les rangs plutôt que sur les données brutes?” [Oui, Non] — Si on répond *Oui*, les calculs se feront après avoir remplacé les valeurs quantitatives brutes par leur rang. Les données semi-quantitatives ne sont pas modifiées par cette procédure, sauf pour ce qui est des valeurs liées (*ex aequo*) qui sont traitées comme en statistique non-paramétrique.

(6) “Fichier de sortie. Intervalle de confiance - Valeur de rejet:” La réponse se donne en pressant l’un des quatre boutons [•0.005 •0.01 •0.05 •0.10] — Le niveau de signification établi ici sert au calcul de la valeur critique de la statistique du périodogramme. La valeur critique apparaît, dans le fichier de résultats, comme une valeur numérique ainsi que comme un symbole “+” dans le graphique.

Les premiers calculs (lecture des données, division en classes) sont effectués à ce point-ci.

(7) “Intervalle de l’analyse: de $\mathbf{x_1}$ à $\mathbf{x_2}$” [OK] — Les périodes incluses dans la fenêtre d’observation d’un périodogramme vont de $T = 2$ à $T = n/2$ où n est le nombre total d'observations dans la série; la valeur affichée pour $\mathbf{x_1}$ est donc 2 alors que la valeur de $\mathbf{x_2}$ est $n/2$. Si la série est longue, l’usager peut désirer une fenêtre de calcul plus étroite; il peut changer les valeurs de $\mathbf{x_1}$ et de $\mathbf{x_2}$ à volonté, avant de presser le bouton *OK*. Cette question n’est posée que dans la version Macintosh du programme. On détermine ainsi le nombre de classes qui seront illustrées dans le périodogramme et incluses éventuellement dans le calcul de la correction de Bonferroni.

Le périodogramme est calculé et apparaît à l’écran. Les périodes significatives sont mises en évidence par différentes teintes de gris correspondant aux niveaux de probabilité suivants:

Niveau de signif.:	p≤0.001	p≤0.01	p≤0.05	p≤0.10	p>0.10
Symbole:	****	***	**	*	

Le menu “Dessin” permet d’imprimer le graphique ou de le conserver dans un fichier de type PICT, ce qui permettra de l’éditer à l’aide d’un programme graphique ou de l’inclure dans un fichier de texte. On peut également demander la “Correction de Bonferroni” à partir de ce même menu si on désire corriger l’effet des tests multiples sur le niveau de signification employé. Cette correction consiste à employer un niveau de signification plus contraignant $\alpha' = \alpha$ / (nombre de tests réalisés simultanément); voir Cooper (1968) ou Miller (1977). Si par exemple on réalise 7 tests simultanés (7 périodes), la correction de Bonferroni modifie le niveau de signification α en $\alpha' = \alpha / 7$, ce qui peut changer la signification de certaines périodes du périodogramme (voir l’exemple). Pour la même

raison, et suivant Oden (1984), nous avons recommandé d'employer la correction de Bonferroni dans le cas des corrélogrammes (programmes AUTOCORRÉLATION SPATIALE et MANTEL).

Il suffit de "Terminer" le graphique pour retourner au menu "Fichiers" qui permet de traiter immédiatement un autre fichier. La commande "Interrompre" dans le menu "R: Period" permet de quitter le programme.

Exemple

L'exemple suivant illustre l'utilisation du programme en version pour grands ordinateurs. La série de données comprend les 16 valeurs semi-quantitatives suivantes:

```
2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 3
```

Le fichier d'appel, dont le dialogue forme la première partie de l'exemple, demande le nom du fichier d'entrée; cet exemple a été réalisé sous CMS.

```
*** Avez-vous verifie
*** ... si toutes les valeurs sont STRICTEMENT positives?
*** ... si les variables forment bien les LIGNES du fichier?
*** ... si le pas entre les donnees est constant?

Quel est le nom du fichier de DONNEES? (Par defaut: "... data a")
semiq  16  a

Execution begins...

 P E R I O D O G R A M M E   D E   C O N T I N G E N C E

 VERSION 2.0b

 UNIVERSITE DE MONTREAL
 DEPARTEMENT DE SCIENCES BIOLOGIQUES
 CASE POSTALE 6128, SUCC. "A"
 MONTREAL, P.Q. H3C 3J7

 AUTEUR:  A. VAUDOR

 NOMBRE D'OBSERVATIONS
16
 NOMBRE DE VARIABLES
1
 LES DONNEES SONT-ELLES DEJA DIVISEES EN CLASSES? (O ou N)
n

 Noter que ce programme emploie la procedure EXACT,  lorsque
 cela  est possible,  pour diviser une variable  en classes.
 Ceci peut conduire a une meilleure partition  que la proce-
 dure pas a pas decrite par Legendre et al. (1981: 969-973),
 procedure qui porte le nom de APPROX dans ce programme.

 NOMBRE DE CLASSES ? (0 POUR CALCUL PAR LE PROGRAMME)
0
 DESIREZ-VOUS TRANSFORMER LES DONNEES EN RANGS ?
```

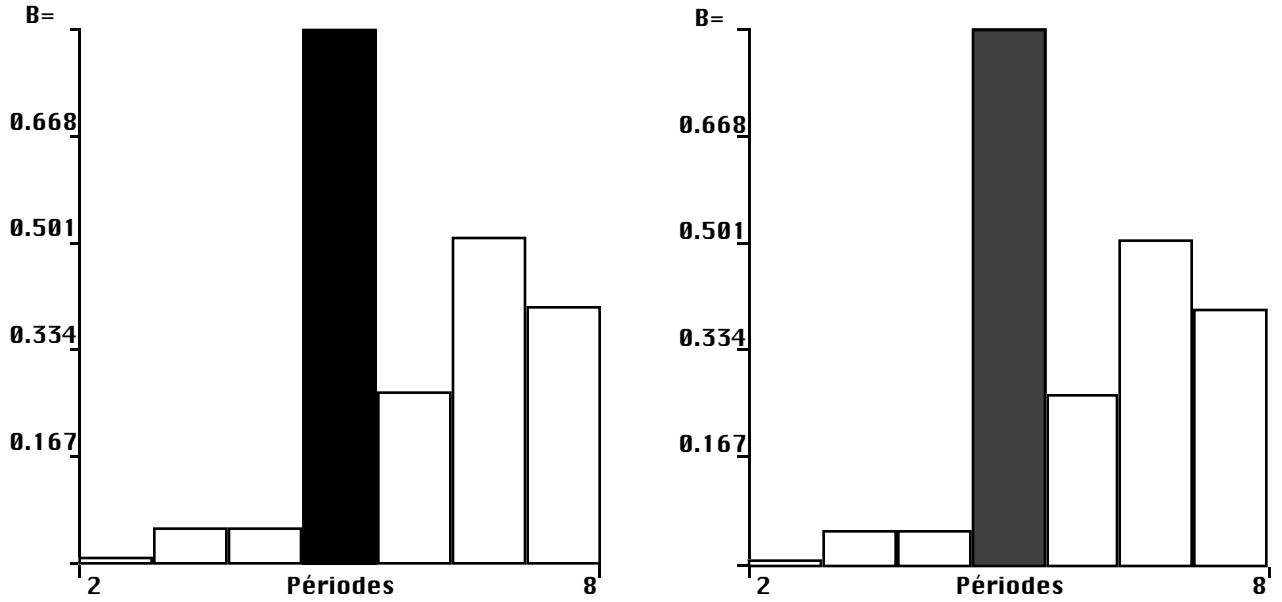
```
 DESIREZ-VOUS TRANSFORMER LES DONNEES EN RANGS ?
o
 CHOIX DE L'INTERVALLE DE CONFIANCE:
 1 pour  0.005 , 2 pour  0.01, 3 pour  0.05, 4 pour  0.10
3

 TABLEAU DE CONTINGENCE:

 NOMBRE DE CLASSES:    3
        CLASSE   BORNE SUPERIEURE
           1         3.00000
           2         6.00000
           3        10.00000

 H(S)/S :    0.52043    EN LOG BASE 2

             P E R I O D O G R A M M E   D E   C O N T I N G E N C E


     (+=INTERVALLE DE CONFIANCE, *==>(+=B) )     L'ECHELLE DE B EST EN LOG. NAT.

 T=_ 0      0.27      0.54      0.81      1.08             VALEUR
    .         .         .         .         .      B      CRITIQUE  PROB(2NB)
   2B      +  .         .         .         .   0.00000   0.18719   1.00000
   3.  B      .+        .         .         .   0.07630   0.29656   0.65514
   4.    B    .    +    .         .         .   0.12912   0.39375   0.65885
   5.         .       + .         B         .   0.82227   0.48437   0.00093 ****
   6.         B         .+        .         .   0.25824   0.57187   0.60311
   7.         .         . B +     .         .   0.58357   0.65625   0.09670 *
   8.         .   B     .      +  .         .   0.38905   0.74062   0.57025
Fin du programme.
```

Graphiques et contenu du fichier de résultats

Les deux graphiques reproduits plus bas sont les périodogrammes de contingence tels qu'ils apparaissent à l'écran du Macintosh. Le fichier suivant, soumis à l'analyse, contient une série de 16 valeurs pour une variable qualitative:

```
1 1 2 3 3 2 1 2 3 2 1 1 2 3 3 1
```

Cet exemple est également analysé dans l'article de Legendre *et al.* (1981). Selon que l'on demande ou non la correction de Bonferroni sur les probabilités, on obtient l'un ou l'autre des deux graphiques suivants:



Sans correction pour tests multiples *Après correction de Bonferroni*

Lorsqu'on applique la correction de Bonferroni, le niveau de signification change. Dans le cas d'espèce, il y a 16 valeurs, donc le programme pourra analyser les périodes de 2 à 8, soit 7 périodes; puisqu'on réalise 7 tests simultanés, la correction de Bonferroni modifie le niveau de signification α en $\alpha' = \alpha / 7$, ce qui change la signification de la probabilité de la période 5:

Niveau de signif. α avant correction	Après correction de Bonferroni: $\alpha' = \alpha / 7$
0.10 *	0.01429 *
0.05 **	0.00714 **
0.01 ***	0.00143 ***
0.001 ****	0.00014 ****

Période	Prob.(H_0)	Signification avant correct.	Signification après correct.
2	0.81762		
3	0.77290		
4	0.94024		
5	0.00079	****	***
6	0.56404		
7	0.17769		
8	0.53819		

Le fichier de sortie Macintosh contient les informations concernant la division des variables quantitatives ou semi-quantitatives en classes, ainsi que des détails additionnels sur le périodogramme de contingence. La liste ci-dessous résulte de l'analyse du fichier suivant, où la variable (16 valeurs) est semi-quantitative (même fichier qu'au paragraphe Exemple ci-dessus):

```
2 2 4 7 10 5 2 5 8 4 1 2 5 9 6 3
```

Cet exemple, qui est également analysé dans l'article de Legendre *et al.* (1981), n'est donc pas le même que celui ayant servi à produire les deux graphiques ci-dessus.

```
PERIOD: Périodogramme de contingence
        (Version 3.0)
```

```
Auteur: A. Vaudor
Département de sciences biologiques, Université de Montréal,
C. P. 6128, succursale A, Montréal, Québec H3C 3J7.

FICHIER DE DONNEES: 16 données quantitatives

Tableau de contingence

Nombre de classes:   3
Classe   Limite
   1    3.00000
   2    6.00000
   3   10.00000

h(s)/s :    0.52043
```

Cette première partie n'est présentée que lorsque le programme a dû diviser une variable ordonnée (quantitative ou semi-quantitative) en classes. La limite **supérieure** de chaque classe est fournie par le programme, de même que la quantité d'entropie par classe pour cette division ["h(s)/s"]. Voir la remarque dans le paragraphe d'introduction à propos de l'algorithme EXACT utilisé dans le programme, par rapport à l'algorithme pas à pas décrit dans l'article de Legendre *et al.* (1981). La liste de sortie se poursuit par le périodogramme de contingence lui-même:

```
      Périodogramme de contingence

 (+=Intervalle de confiance, *==>(+=B) )    Echelle en log. nat.

  T=\ 0      0.27      0.54      0.81      1.08 Valeur
    .         .         .         .         .   B         critique  prob(2nb)
   2B    +    .         .         .         .   0.00000   0.14406   1.00000
   3.  B     +.         .         .         .   0.07630   0.24313   0.65514
   4.    B    . +       .         .         .   0.12912   0.33125   0.65885
   5.         .    +    .         B         .   0.82227   0.41875   0.00093 ****
   6.         B       + .         .         .   0.25824   0.50000   0.60311
   7.         .         .+B       .         .   0.58357   0.57812   0.09670 *
   8.         .   B     .   +     .         .   0.38905   0.65938   0.57025
```

Ce graphique représente un périodogramme dont l'abscisse (périodes *T*) va du haut vers le bas alors que l'ordonnée (entropie commune *B* calculée en logarithmes naturels) va de la gauche vers la droite. Le symbole "B" est employé dans le graphique pour représenter les valeurs de la statistique *B*. La valeur critique, pour la probabilité fournie en réponse à la question (6) (sans correction de Bonferroni), est représentée par des "+"; la probabilité demandée pour l'intervalle de confiance est ici de 0.1. Les trois colonnes de nombres fournissent la valeur précise de la statistique *B*, la valeur critique au seuil de probabilité prédéterminé ainsi que la probabilité de l'hypothèse nulle (probabilité que cette valeur de *B* ne soit pas différente de zéro). Enfin, une dernière colonne met en évidence les valeurs significatives aux seuils de signification de 0.10 (*), 0.05 (**), 0.01 (***) ou 0.001 (****), avant que la correction de Bonferroni ne soit appliquée. Lorsque la série est suffisamment longue, il ne faut pas se surprendre de voir apparaître comme significatives les multiples des périodes de base du phénomène.

PNCOMP Macintosh

Que fait PNCOMP ?

Ce programme produit une ordination en espace réduit par la méthode des composantes principales décrite dans tous les ouvrages de statistiques multidimensionnelles. Cette méthode très générale d'analyse comporte de nombreuses variantes; les principales sont brièvement discutées au tableau 3.

Tableau 3 — Questions que l'on se pose à propos de l'analyse en composantes principales (adapté du tableau 9.I de Legendre & Legendre, 1984a).

Avant de faire une analyse en composantes principales:

1) Les descripteurs sont-ils appropriés?

   → Descripteurs quantitatifs; normalité; pas trop de zéros;
   en principe, plus d'objets que de descripteurs

2) Les descripteurs sont-ils dimensionnellement homogènes? Question 7

   → Si oui: ACP sur la matrice de dispersion (variances-covariances)
   → Si non: ACP sur la matrice de corrélations

3) But visé par l'ordination en espace réduit: Question 8

   → Représenter la position relative des objets: normalisation des vecteurs propres à 1
   → Représenter la corrélation entre les descripteurs: normalisation des vecteurs propres à $\sqrt{\lambda}$
   → Représenter à la fois les objets et les descripteurs: double projection (normalisation à 1)

En examinant les résultats d'une analyse en composantes principales:

1) Quelles sont les valeurs propres significatives? Voir description du fichier de résultats

   → Test: λ_i est-il plus grand que la moyenne des λ?
   → Test: le % de variance de $\hat{\lambda}_i$ est-il plus grand que prévu par le modèle du bâton brisé?

2) Quels sont les descripteurs qui contribuent davantage à la formation de l'espace réduit?

   → Voir le graphique des descripteurs, dans lequel les variables sont représentées par des axes (flèches), ou encore le tableau des coordonnées des descripteurs Question 14
   → Voir les descripteurs qui dépassent le cercle de contribution équilibrée
   → Examiner également les corrélations entre les descripteurs et les axes principaux

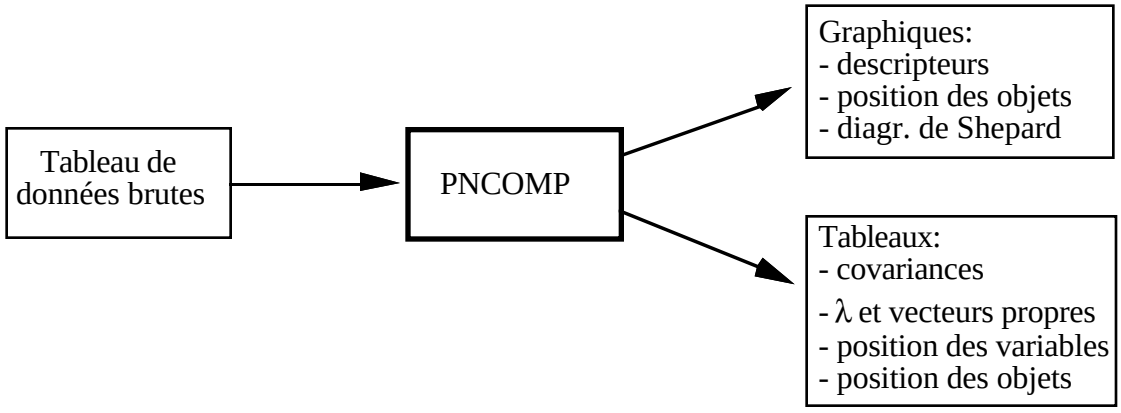
3) Comment trouver la position des objets dans l'espace réduit?

   → Voir le graphique et le tableau des coord. des objets dans l'espace réduit Question 15

4) Les distances entre objets sont-elles bien préservées dans l'espace réduit?

   → Voir le diagramme de Shepard (comparaison des distances) Question 16

Il n'existe qu'une version Macintosh de ce programme, de nombreux logiciels statistiques permettant de réaliser cette analyse sur les grands ordinateurs. Les calculs sont effectués à partir d'une matrice de données brutes qui peut contenir des informations absentes. Le programme produit des graphiques ainsi qu'un fichier de résultats, si l'usager en fait la demande. Le programme ne peut pour le moment traiter plus de 55 variables **[revoir]**.



Fichiers d'entrée et de sortie

(1) Fichier de données brutes

Le fichier de données brutes est un tableau rectangulaire (lignes = objets, colonnes = descripteurs) de données quantitatives, écrit en ASCII sans aucun identificateur de ligne ou de colonne. Ce tableau est souvent extrait d'un chiffrier (option: texte seulement) comportant davantage de lignes et/ou de colonnes, dans lequel ces renseignements sont consignés. Les nombres peuvent être séparés par des espaces, des tabulateurs, etc. et n'ont pas besoin de suivre un format régulier (colonnes bien alignées). Ce tableau peut comporter des absences d'informations, codées par une valeur numérique (par exemple -9, ou -999, etc.) ne portant pas à confusion avec d'autres valeurs présentes dans le tableau. Le tableau peut également comporter des objets supplémentaires (en fin de liste), ainsi que des variables supplémentaires (après les variables actives), qui seront positionnés dans l'espace réduit sans avoir été inclus dans le calcul des valeurs et des vecteurs propres. Enfin, si les objets appartiennent à des groupes identifiés au préalable, une variable (nombres entiers positifs) décrivant cette appartenance peut être incluse dans le tableau des données, ce qui permettra au programme d'identifier les groupes d'objets par des symboles différents dans le graphique; cette variable peut être située n'importe où parmi les colonnes du fichier.

(2) Fichier de résultats

Le fichier de résultats contiendra les tableaux que l'usager aura demandé d'y inscrire à partir du menu "Calculs (détails)", à savoir: le tableau des covariances ou des corrélations, les valeurs et vecteurs propres, la position des variables et la position des objets par rapport aux premières composantes principales.

Les options du programme

Ce programme permet de réaliser les calculs des valeurs et des vecteurs propres à partir soit de la matrice des covariances, soit de la matrice des corrélations (qui sont les covariances des données centrées réduites). Les vecteurs propres peuvent être normés à la longueur 1 (si on est intéressé avant tout à exprimer dans l'espace réduit les relations de distance euclidienne entre les objets) ou encore à la racine carrée de leur valeur propre (si on est davantage intéressé à exprimer les corrélations entre descripteurs). Deux types de rotations sont également disponibles. Des objets ou des variables

supplémentaires peuvent être projetés dans l'espace réduit, suivant en cela la tradition de l'école française d'analyse des données.

S'il y a des absences d'information dans le tableau des données, deux stratégies sont disponibles dans ce programme. D'une part, les objets porteurs de telles informations absentes peuvent être simplement éliminés de l'analyse (*"listwise deletion of missing values"*). D'autre part, lors du calcul des covariances ou des corrélations, toute paire impliquant une absence d'information peut être éliminée des calculs (*"pairwise deletion of missing values"*); ceci donne naissance à des covariances possédant un nombre inégal de degrés de liberté, ce qui permet éventuellement l'apparition de petites valeurs propres négatives qui devront être négligées lors de l'interprétation (voir aussi la discussion des valeurs propres négatives, à la section du programme PCOORD). La solution qui consiste à estimer les valeurs absentes n'est pas disponible pour le moment dans *R*.

Les questions du programme

Les questions présentées par le programme à l'écran du Macintosh sont décrites dans les paragraphes qui suivent. Pour faire démarrer le programme, il faut cliquer sur l'icône, puis donner la commande "Ouvrir" dans le menu "Fichier".

(1) "Nombre d'objets (lignes), à l'exception des objets supplémentaires" — On inscrit le nombre d'objets qui seront inclus dans le calcul des valeurs et des vecteurs propres, y compris les objets porteurs d'informations absentes.

(2) "Nombre de variables (colonnes), à l'exception des variables supplémentaires et de la variable identifiant les groupes" — On inscrit le nombre de variables qui seront incluses dans le calcul des valeurs et des vecteurs propres.

(3) "Nombre d'objets supplémentaires" — Les objets supplémentaires, qui seront positionnés dans le graphique-objets sans toutefois avoir été inclus dans le calcul des valeurs et des vecteurs propres, doivent occuper les dernières lignes du tableau.

(4) "Nombre de variables supplémentaires" — Les variables supplémentaires doivent occuper des colonnes situées plus à droite que les colonnes portant les variables incluses dans le calcul des valeurs et des vecteurs propres.

(5) "Y a-t-il de l'absence d'information?" [Oui, Non] — Voir la description des méthodes d'exclusion des informations absentes, au dernier paragraphe de la section précédente (Options du programme).

(5.1) "Supprimer tous les objets contenant cette absence?" [Oui, Non] — Si on répond *Oui*, la première méthode est employée (suppression des objets porteurs d'informations manquantes). Si on répond *Non*, c'est la seconde méthode qui sera employée (calcul de covariances ou de corrélations basées sur un nombre inégal de paires d'objets).

(5.2) "Valeur indiquant l'absence d'information" — On inscrit quelle **valeur numérique** a été utilisée dans le fichier pour indiquer qu'une information est absente (souvent: -1, -9, -999, etc.)

(6) "Fichier de données" — Le programme présente le menu des fichiers ASCII disponibles.

(7) "Calculs sur la matrice de corrélations plutôt que sur les covariances?" [Oui, Non] — On n'effectue les calculs à partir de la matrice de corrélations (qui sont les covariances des variables centrées réduites) que lorsque les variables ne sont pas de même nature ou ne sont pas dimensionnellement homogènes (mesurées dans les mêmes unités physiques); la réduction (division par l'écart type) élimine les effets des échelles de mesure en produisant des variables sans dimensions

physiques. Lorsque les descripteurs sont de même nature et mesurés dans les mêmes unités, c'est la matrice de dispersion qu'il faut employer comme base des calculs. Les composantes principales extraites de la matrice de corrélations ne sont pas les mêmes que celles extraites de la matrice de dispersion.

(8) "Normalisation par les √(lambdas) ?" [Oui, Non] — La normalisation des vecteurs propres à 1 préserve la distance euclidienne entre les objets, dans l'espace de pleine dimension; les axes d'origine demeurent orthogonaux lors de cette normalisation. La représentation en espace réduit produit donc une projection du nuage de points d'origine en quelques dimensions. Par ailleurs, la normalisation à la racine carrée de la valeur propre ($\sqrt{\lambda}$) fait en sorte que les axes-descripteurs forment entre eux un angle proportionnel à leur covariance. L'angle varie de 0° (covariance positive maximale) à 180° (covariance négative maximale), un angle de 90° signifiant une covariance nulle. On utilise donc cette normalisation lorsque le but de l'analyse est de représenter les relations entre descripteurs par des projections angulaires. Les relations de distance entre les points sont déformées lors de cette transformation.

(9) "Combien de valeurs propres à extraire?" — L'algorithme utilisé pour le calcul des valeurs propres est un algorithme pas à pas **[nom?]** qui calcule d'abord les valeurs propres les plus importantes. L'usager peut limiter le calcul aux quelques premières valeurs propres (habituellement de 2 à 5) qui contiennent habituellement la plus grande partie de la variance; cela peut représenter un gain de temps appréciable pour les analyses comportant de nombreux descripteurs.

(10) "Le fichier d'entrée contient-il des identificateurs de groupes?" [Oui, Non] — Si les objets appartiennent à des groupes identifiés au préalable, une variable (nombres entiers positifs) décrivant cette appartenance peut être incluse dans le tableau des données, ce qui permettra au programme d'identifier les groupes d'objets par des symboles différents dans le graphique. Cette variable peut être située n'importe où parmi les colonnes du fichier; sa position est précisée à la question (10.1).

(10.1) "Numéro de la variable identifiant les groupes d'objets" — Cette variable, dans laquelle les numéros de groupes sont codés par des **entiers positifs**, peut être située n'importe où dans le fichier. On indique ici quelle colonne elle occupe. Si la colonne désignée contient autre chose que des entiers positifs, le programme émet un message d'erreur et s'arrête.

Le fichier de données est lu à ce point-ci.

(11) "Titre de ce travail ..." — On fournit un titre, qui sera repris en en-tête des graphiques reproduits par l'imprimante.

Les valeurs propres et les vecteurs propres sont calculés à ce point-ci.

(12) "Rotation Varimax?" [Oui, Non] — La rotation Varimax normalisée (Kaiser, 1958) est une rotation orthogonale du nuage de points qui tente de simplifier les colonnes du tableau des vecteurs propres (normalisés au préalable à 1) en maximisant la variance du carré des saturations de chaque colonne; lorsque la variance des saturations est grande, celles-ci ont tendance à être près de 0 ou de 1. La rotation Varimax maximise la somme de ces variances pour tous les facteurs soumis à la rotation. Des groupes d'axes-descripteurs ont ainsi plus de chance de se trouver près (*i.e.*, à angle faible) des axes factoriels après rotation, ce qui simplifie l'interprétation de ces facteurs en termes des variables d'origine. La quantité de variance expliquée par un sous-espace factoriel demeure inchangée après rotation. Les facteurs demeurent non corrélés après cette rotation orthogonale. La rotation est réalisée pour le nombre d'axes principaux que l'usager aura indiqué en réponse à la question (9).

(13) "Rotation de Harris-Kaiser?" [Oui, Non] — La rotation de Harris & Kaiser (1964), appelée aussi *orthoblique*, introduit une déformation des angles entre axes-descripteurs. La rotation procède en trois étapes: (1) déformation des vecteurs propres, dont l'intensité est déterminée à la question (13.1); (2) rotation Varimax; (3) déformation inverse de celle de l'étape 1. Les facteurs deviennent corrélés après

cette rotation oblique.

(13.1) “Coefficient de déformation de l’espace” — La déformation de l’espace est déterminée en spécifiant l’exposant à donner à la racine carrée des valeurs propres. Les valeurs peuvent aller de 0 à 1, la valeur 1 correspondant à la solution Varimax. Ce coefficient est le même que le paramètre HKPOWER de la procédure FACTOR de SAS.

(14) “Graphique des descripteurs?” [Oui, Non] — Ces graphiques montrent la projection des axes-descripteurs dans l’espace réduit; les axes-descripteurs y sont donc représentés comme des axes.

(14.1) “Numérotation des variables sur le graphique?” [Oui, Non] — Si on répond *Oui*, des numéros séquentiels permettent d’identifier les variables sur chacun des graphiques.

(14.2) “Nombre de dimensions à représenter?” — Des graphiques successifs seront produits pour toutes les paires d’axes principaux demandés. Ainsi, si on demande de représenter 3 dimensions, trois graphiques seront produits, correspondant respectivement aux axes I et II, I et III, II et III.

Les graphiques des descripteurs sont produits à ce point-ci. Pour augmenter la résolution, on peut agrandir n’importe quelle partie de l’image en l’entourant d’un cadre à l’aide de la souris. Si on a choisi de réaliser les calculs sur la matrice de corrélations, ou encore sur la matrice de covariances avec une normalisation des vecteurs propres à une longueur de 1 à la question 8, le cercle de contribution équilibrée apparaît également sur les graphiques. Legendre & Legendre (1984a) ont montré que si tous les n descripteurs contribuaient de façon égale à la formation de l’espace réduit en d dimensions (d étant le nombre de dimensions choisi en réponse à la question 14.2), alors chacun d’eux aurait une longueur de $\sqrt{d/n}$. Par conséquent, si on trace un cercle de rayon égal à $\sqrt{d/n}$, alors tout axe-descripteur qui dépasse ce cercle contribue davantage à l’espace réduit que ne le prédit le modèle de la contribution équilibrée des descripteurs. Lorsque les calculs sont réalisés à partir de la matrice de covariances et que les vecteurs propres sont normés à $\sqrt{\lambda}$, la formule de calcul des contributions équilibrées est un peu plus complexe et ne donne plus naissance à un cercle (Legendre & Legendre, 1984a); c’est pourquoi le cercle de contribution équilibrée n’est pas tracé dans ce cas.

Les tableaux suivants sont disponibles dans le menu “Calculs (détails)”: le tableau des covariances ou des corrélations, les valeurs et vecteurs propres, ainsi que la position des variables par rapport aux composantes principales sélectionnées en réponse à la question (9). On peut monter ou descendre dans ces tableaux en pointant le curseur de la souris dans le bas ou le haut du tableau. Ces résultats peuvent être envoyés directement à l’imprimante ou copiés dans un fichier de résultats pour référence future. De même, à partir du menu “Graphiques”, les graphiques peuvent être envoyés à l’imprimante, ou encore on peut les préserver dans un fichier de type PICT, ce qui permettra de les éditer à l’aide d’un programme graphique ou de les inclure dans un fichier de texte. Il faut “Terminer” chaque graphique pour passer au suivant, ou encore pour passer à la question suivante.

(15) “Graphique des objets?” [Oui, Non] — Ces graphiques montrent la projection des objets dans l’espace réduit; les objets y sont donc représentés par des points.

(15.1) “Numérotation des objets sur le graphique?” [Oui, Non] — Si on répond *Oui*, des numéros séquentiels permettent d’identifier les objets sur chacun des graphiques.

(15.2) “Nombre de dimensions à représenter?” — Des graphiques successifs seront produits pour toutes les paires d’axes principaux. Ainsi, si on demande de représenter 3 dimensions, trois graphiques seront produits, correspondant respectivement aux axes I et II, I et III, II et III.

Les graphiques des objets sont produits à ce point-ci. Pour augmenter la résolution, on peut agrandir n’importe quelle partie de l’image en l’entourant d’un cadre à l’aide de la souris. La liste des “Positions des objets” par rapport aux composantes principales sélectionnées en réponse à la question (9) devient maintenant disponible dans le menu “Détails de calcul”. Il faut “Terminer” chaque

(9) devient maintenant disponible dans le menu “Détails de calcul”. Il faut “Terminer” chaque graphique pour passer au suivant, ou encore pour passer à la question suivante.

(16) “Comparaison des distances (Diagramme de Shepard)?” [Oui, Non] — Cette question n’apparaît que si on a choisi de faire les calculs à partir de la matrice des covariances. Si on y répond *Oui*, les questions suivantes permettent de préciser comment se fera la comparaison entre les distances d’origine (distances euclidiennes entre les objets, calculées à partir du fichier de données brutes) et les distances dans l’espace réduit à 2, 3, … dimensions. Dans ce graphique, un nuage de points étroit, situé sous la diagonale mais près de celle-ci, indique une bonne représentation des distances d’origine dans l’espace réduit. Occasionnellement, des points pourront apparaître au-dessus de la diagonale; lorsque cela se produit, ces points correspondent à des objets pour lesquels des informations absentes ont été remplacées par le programme (voir la question 5.1).

(16.1) “Diagramme de Shepard: combien de vecteurs propres?” — On indique combien de dimensions de l’espace réduit seront incluses dans cette comparaison des distances (en général, 2 ou 3).

(16.2) “XX distances à calculer; préférez-vous échantillonner celles-ci?” [Oui, Non] — Il y a XX = $p(p\text{-}1)/2$ distances entre p objets. Lorsque ce nombre devient trop grand (plus de quelques centaines: calcul trop long), l’usager peut demander à l’ordinateur de choisir au hasard un certain nombre de ces distances. Le nombre en question sera déterminé à la question (16.3), la sélection étant réalisée au hasard à l’aide d’un générateur de nombres pseudo-aléatoires initialisé à la question (16.4).

(16.3) “Nombre de distances à échantillonner” — On inscrit le nombre désiré.

(16.4) “Générateur de nombres aléatoires: entrer un (petit) chiffre” — On inscrit un petit entier positif, par exemple 2, 5 ou 10.

(16.5) “Autre comparaison des distances?” [Oui, Non] — Si on répond *Oui* à cette question, on retourne à la question (16.1).

(17) “Calculs terminés?” [Oui, Non] — On répond *Non* si on désire effectuer une rotation, par exemple; dans ce cas, les questions (12) à (16) sont présentées à nouveau. La réponse *Oui* provoque la fin du programme.

Exemple

L’exemple ci-dessous présente une analyse en composantes principales d’un fichier de données physico-chimiques portant sur 71 stations d’échantillonnage en milieu aquatique; 11 variables ont été mesurées. Comme celles-ci sont exprimées dans des unités physiques différentes (mg/L, °C, etc.), il convient de réaliser l’analyse à partir de la matrice des corrélations entre descripteurs. Une douzième variable décrit l’appartenance des observations à l’un ou l’autre de 6 groupes, qui seront représentés par différents symboles dans les graphiques. On a demandé le calcul de 3 valeurs propres.

Graphiques et contenu du fichier de résultats

Le fichier de résultats peut contenir l’un ou l’autre des tableaux que l’on aura demandé d’y inscrire, à savoir: le tableau des covariances ou des corrélations, les valeurs et vecteurs propres, la position des variables ainsi que la position des objets par rapport aux composantes principales sélectionnées en réponse à la question (9). Puisqu’il est écrit en ASCII, ce fichier peut aisément être édité si on désire transférer certains de ces résultats à un autre programme. Un exemple de diagramme de Shepard est présenté avec le fichier de résultats du programme PCOORD.

```
Matrice de corrélations

          1         2         3         4         5         6         7
 1    1.0000
 2    0.2861    1.0000
 3   -0.2737   -0.0784    1.0000
 4   -0.4857   -0.8501    0.0283    1.0000
 5   -0.4207   -0.6456   -0.1422    0.8441    1.0000
 6   -0.0502   -0.0926    0.2120   -0.0019   -0.0187    1.0000
 7    0.0115   -0.2906    0.1854    0.4446    0.3940    0.0088    1.0000
 8   -0.6607   -0.4657    0.0515    0.7033    0.7368    0.0875    0.2165
 9   -0.3879   -0.1814   -0.1466    0.3345    0.4184    0.2141   -0.1736
10    0.2577    0.5017    0.0006   -0.6069   -0.5282    0.0195   -0.3897
11    0.2701    0.4822    0.0529   -0.5764   -0.5810    0.0861   -0.1803

          8         9        10        11
 8    1.0000
 9    0.4824    1.0000
10   -0.4594   -0.1870    1.0000
11   -0.6025   -0.2748    0.6542    1.0000

Valeurs et vecteurs propres

Moyenne des valeurs propres :   1.00000 ON PEUT INTERPRETER LES LAMBDAS
PLUS GRANDS QUE CETTE VALEUR (Ref.: Ecologie numérique T.2, P. 123)

        VALEURS PROPRES     % DE VARIANCE        % BATON BRISE
            4.72621            42.96551            27.45343
            1.49324            13.57489            18.36252
            1.36044            12.36767            13.81707
```

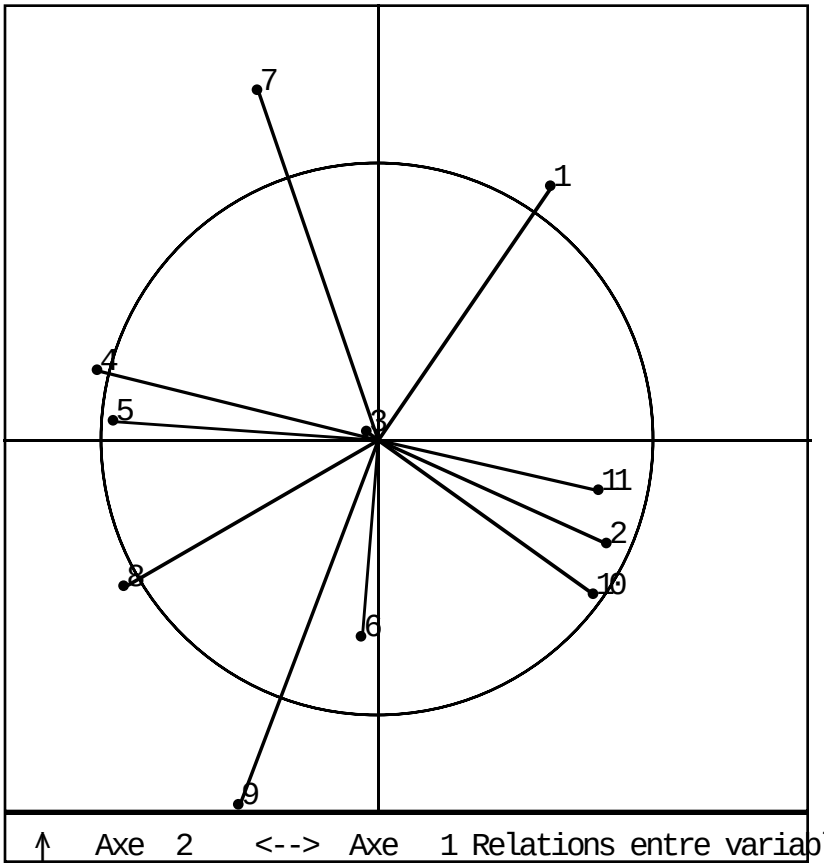
Les critères suivants peuvent aider à déterminer combien de valeurs propres il faut retenir. D'une part, on peut décider de ne retenir que les valeurs propres qui sont plus grandes que la moyenne des λ, puisqu'on peut démontrer qu'une variable issue d'un générateur de nombres pseudo-aléatoires deviendrait dominante à partir de cette valeur propre; notez que lorsque les calculs sont réalisés à partir de la matrice de corrélations, comme c'est le cas ici, la moyenne des valeurs propres est égale à 1. D'autre part, on peut comparer le pourcentage de variance expliqué par les valeurs propres successives à la distribution aléatoire du bâton brisé (voir Frontier, 1976, ou encore "Écologie numérique, tome 2, page 124). Toute valeur propre qui explique davantage de variance que la fraction correspondante du modèle aléatoire du bâton brisé vaut la peine d'être examinée.

```
VECTEURS PROPRES SELECTIONNES (PAR COLONNES, NORME = 1)

 1             0.26817             0.38985            -0.22411
 2             0.35148            -0.15842            -0.07806
 3            -0.01744             0.01350             0.75352
 4            -0.42825             0.10752             0.02143
 5            -0.40733             0.02935            -0.13672
 6            -0.02329            -0.30288             0.46169
 7            -0.18527             0.53790             0.26671
 8            -0.38760            -0.22506             0.02097
 9            -0.21302            -0.56205            -0.17144
10             0.33065            -0.23811             0.05607
11             0.33929            -0.07738             0.19866
POSITION DES OBJETS DANS LE NOUVEL ESPACE
```

```
POSITION DES OBJETS DANS LE NOUVEL ESPACE
          1         2         3
   1    2.2939   -0.7890    0.5123
   2    2.2939   -0.7890    0.5123
   3    2.5417   -0.6558    0.8178
[etc.]
  70   -0.2559    1.5296   -0.9209
  71   -0.2559    1.5296   -0.9209
```

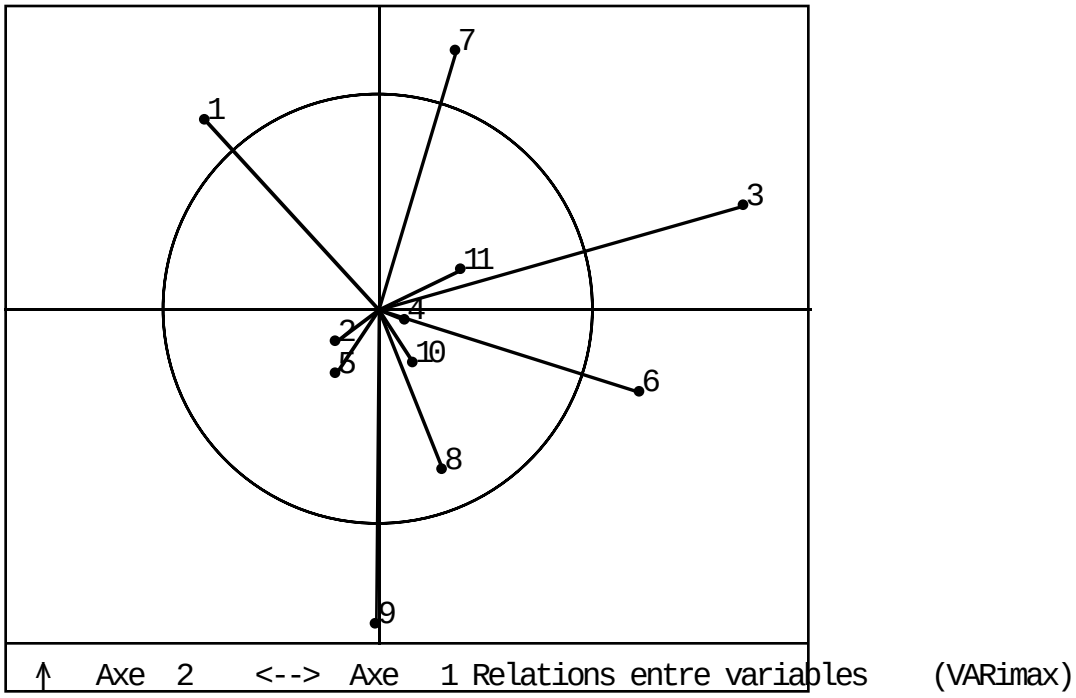
```
POSITION DES VARIABLES DANS LE NOUVEL ESPACE (Méthode (VARimax))
          1         2         3
   1   -0.3435    0.3767    0.1193
   2   -0.0835   -0.0635    0.3791
   3    0.7243    0.2081    0.0174
   4    0.0511   -0.0208   -0.4386
   5   -0.0852   -0.1286   -0.4021
   6    0.5194   -0.1619    0.0972
   7    0.1531    0.5109   -0.3321
   8    0.1244   -0.3145   -0.2949
   9   -0.0036   -0.6242   -0.0327
  10    0.0673   -0.1070    0.3914
  11    0.1652    0.0811    0.3560
```

Les graphiques suivants sont produits à la demande de l'usager. Voici d'abord le graphique de la projection des descripteurs dans l'espace réduit, sans rotation:

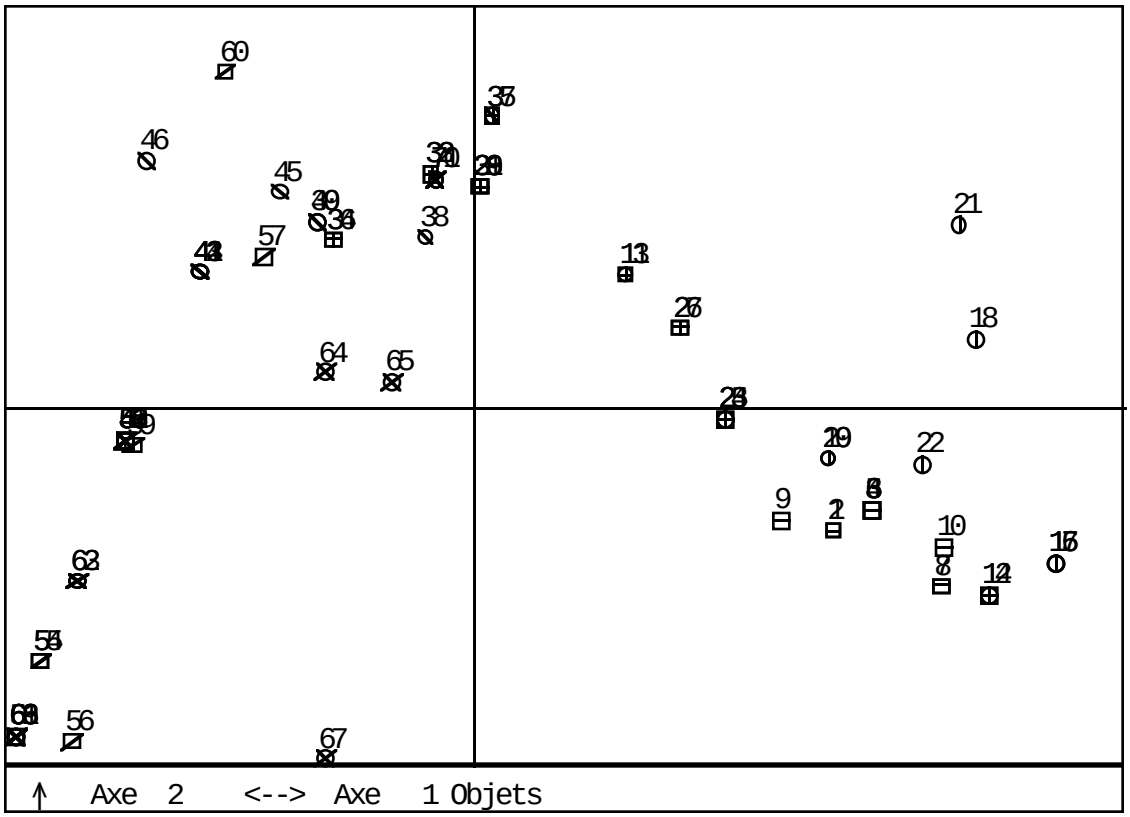


Le second graphique représente les descripteurs dans l'espace réduit, après une rotation Varimax

sur les trois dimensions de l'espace factoriel:



Le troisième graphique montre la position des objets dans l'espace réduit des deux premières composantes principales. Notez les symboles identifiant les groupes d'objets.

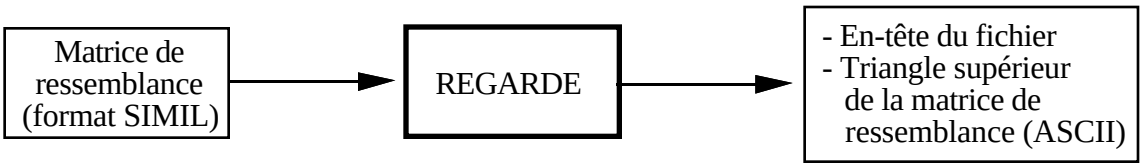


REGARDE

Que fait REGARDE ?

Ce programme permet de voir en clair le contenu des fichiers binaires produits par SIMIL, IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh); ces matrices, qui sont écrites en binaire et non en ASCII, ne peuvent être lues directement. Comme l'usager n'est souvent intéressé qu'à examiner le contenu du bloc d'informations générales du fichier (titre, date de fabrication, fonction, nombre d'objets et nombre de descripteurs), une première fenêtre, dans la version Macintosh, présente uniquement ces informations; ensuite, l'usager est invité à demander la transcription du fichier binaire sur un fichier ASCII, s'il le désire. Dans le fichier de sortie, après le bloc d'identification du fichier, seul le triangle supérieur de la matrice de ressemblance est présenté, accompagné des noms d'objets identifiant les lignes et les colonnes.

Fichiers d'entrée et de sortie



(1) Le fichier d'entrée

Le fichier d'entrée est un fichier binaire de similarités, de distances ou de mesures de dépendance entre descripteurs, écrit par les programmes SIMIL, IMPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh). Dans la version Macintosh, ces fichiers sont représentés par une icône triangulaire portant le nom SIMIL.

(2) Le fichier de sortie

Le fichier de sortie contient deux types d'informations: d'abord, le bloc d'informations générales du fichier (titre, date de fabrication, fonction, nombre d'objets et nombre de descripteurs), qui est suivi par la matrice triangulaire supérieure des mesures de ressemblance. La diagonale n'est pas écrite; selon qu'il s'agit d'une similarité, d'une distance ou d'un coefficient de dépendance entre descripteurs, la diagonale prend implicitement la valeur 0 ou 1. Les identificateurs d'objets sont écrits à gauche et au haut de la demi-matrice; si aucun identificateur n'a été fourni lors de la création de la matrice de ressemblance, ceux-ci sont remplacés par des numéros séquentiels.

Si on désire une sortie ASCII de toute la matrice (carrée), et non seulement sa partie supérieure, en vue de traitements ultérieurs, il faut employer plutôt le programme EXPORT (versions CMS et VMS) ou IMPORT-EXPORT (version Macintosh).

Les questions du programme

En versions CMS et VMS, les questions du programme se rapportent seulement aux noms des fichiers d'entrée et de sortie.

Après présentation de l'en-tête à l'écran, la version Macintosh du programme demande si l'usager désire recopier la matrice de ressemblance sur le fichier de sortie; dans bien des cas, en effet, on ne désire que consulter l'en-tête. Après avoir traité un premier fichier, le programme demande s'il y a un "Autre fichier à traiter ?" Si c'est le cas, on choisit le fichier dans le menu. Il suffit de presser le bouton "Cancel" pour indiquer qu'il n'y a plus de fichiers à traiter; ceci entraîne la fin du programme.

Exemple

Voici un exemple d'utilisation du programme sur grands ordinateurs; cet exemple a été réalisé sous CMS.

```
Quel est le nom du fichier de type SIMIL?
(Par defaut: "... data a")
localite  d1  a

Quel nom doit recevoir le fichier de sortie de ce programme,
contenant la matrice traduite en clair? (Par defaut: "MAT data a")
localite  sortie  a

Execution begins...
```

Contenu du fichier de résultats

Voici un exemple du fichier de sortie. On y trouve d'abord le bloc d'informations concernant le fichier d'entrée (titre, date de fabrication, fonction, nombre d'objets et nombre de descripteurs), suivi de la demi-matrice de ressemblance.

```
P R O G R A M M E   R E G A R D E
pour examiner une matrice produite par SIMIL

VERSION M1.2
Auteur: A. Vaudor

FICHIER D'ENTREE : 7 localités/D1
    TITRE: 7 localités du Québec / Distances en km
    DATE: 7/9/91
    FONCTION: D01
    Nombre d'objets : 7
    Nombre de descripteurs : 2
```

	Chicoutimi	Gaspé	Miguasha	Montréal	Québec	Rimouski
Chicoutimi		485.60595	349.75490	378.53672	180.15916	185.63113	368.84501
Gaspé			161.00831	781.74651	551.23544	302.05421	792.79199
Miguasha				621.47124	392.58719	165.79426	633.74548
Montréal					233.77626	502.99525	44.29501
Québec						269.67142	241.57762
Rimouski							506.65426

SIMIL

Que fait SIMIL ?

SIMIL est un programme de calcul de mesures de ressemblance, calculant des coefficients pour données binaires (présence-absence) ou pour données quantitatives. Ce programme permet le calcul de toutes les mesures exposées au chapitre 7 du manuel de Legendre & Legendre (1984a), à l'exception des coefficients de corrélation partielle. Le tableau 4 donne la liste des coefficients disponibles, alors que les tableaux 5 à 7 résument les critères qui doivent guider l'usager dans le choix d'un coefficient. Quatre types de fichiers, dont le rôle est expliqué en détail à la section suivante, peuvent être utilisés en conjonction avec ce programme; les flèches en tirets indiquent des fichiers qui n'interviennent que pour certains coefficients.

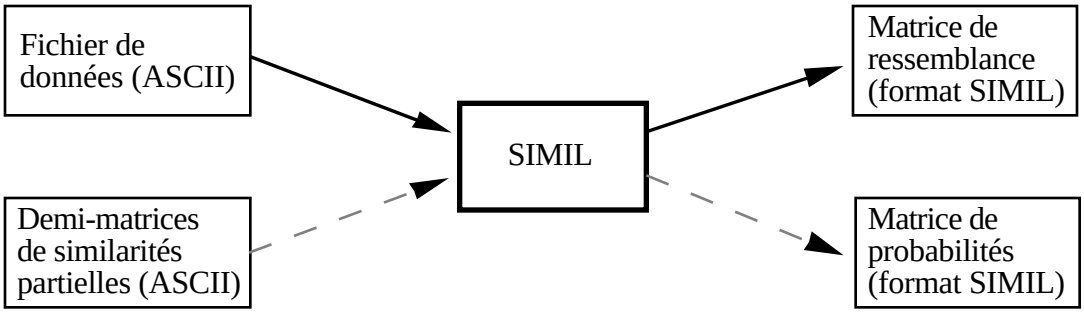


Tableau 4 - Les coefficients d'association du programme SIMIL. Le code reconnu par le programme pour chaque coefficient se trouve dans la colonne de gauche. Les coefficients symétriques incluent les doubles zéros dans la mesure de la ressemblance alors que les coefficients asymétriques n'en tiennent pas compte.

Coefficients binaires incluant les doubles zéros (symétriques)		
S01	(a+d)/(a+b+c+d)	Coefficient de simple concordance (Sokal & Michener)
S02	(a+d)/(a+2b+2c+d)	(Rogers & Tanimoto)
S03	(2a+2d)/(2a+b+c+2d)	
S04	(a+d)/(b+c)	
S05	(1/4) [a/(a+b) + a/(a+c) + d/(b+d) + d/(c+d)]	
S06	ad/√[(a+b)(a+c)(b+d)(c+d)]	
Coefficients binaires excluant les doubles zéros (asymétriques)		
S07	a/(a+b+c)	Coefficient de communauté (Jaccard)
S08	2a/(2a+b+c)	(Sørensen, Dice)
S09	3a/(3a+b+c)	
S10	a/(a+2b+2c)	
S11	a/(a+b+c+d)	(Russell & Rao)
S12	a/(b+c)	(Kulczynski)
S13	(1/2) [a/(a+b) + a/(a+c)]	(Kulczynski)
S14	a/√[(a+b)(a+c)]	(Ochiai)
S26	[a + (d/2)]/(a+b+c+d)	(Faith)

Tableau 4 ***(suite)***

Coefficients quantitatifs incluant les doubles zéros (symétriques)		
S15	Σ(w[i] s[i]) / Σ(w[i])	(Gower, symétrique)
S16	Σ(w[i] s'[i]) / Σ(w[i])	(Estabrook & Rogers)
Coefficients quantitatifs excluant les doubles zéros (asymétriques)		
S17	2W/(A+B)	(Steinhaus)
S18	(1/2) [(W/A) + (W/B)]	(Kulczynski)
S19	Σ(w[i] s[i]) / Σ(w[i])	(Gower, asymétrique)
S20	Σ(w[i] s'[i]) / Σ(w[i])	(Legendre & Chodorowski)
S21		Similarité du khi carré (Roux & Reyssac)
Coefficients probabilistes		
S22		Similarité probabiliste du khi carré
S23		Coefficient probabiliste de Goodall
Coefficients binaires pour l'analyse en mode R (associations d'espèces, etc.)		
S24	[a/√((a+b)(a+c))] - 0.5√(a+c)	(Fager & McGowan)
S25	1 - *p*(khi carré)	(Krylov)
Coefficient de similarité génétique		
NEI		Similarité génétique de Nei (bornée entre 0 et 1)
Coefficients de distance		
D01		Distance euclidienne
D02		Distance moyenne (taxonomique)
D03		Mesure de corde
D04		Métrique géodésique
D05		Distance généralisée de Mahalanobis (entre groupes)
D06		Métrique de Minkowski (l'usager spécifie la puissance)
D07		Métrique de Manhattan
D08		Différence moyenne des descripteurs (Czekanowski)
D09		Indice d'association (Whittaker)
D10		Métrique de Canberra (Lance & Williams)
D11		Coefficient de divergence (Clark)
D12		Coefficient de ressemblance raciale (entre groupes; Pearson)
D13		Coefficient non-métrique (Watson, Williams & Lance)
D14		Différence de pourcentages (Odum; Bray & Curtis)
Coefficients de dépendance (mode R)		
RP		*r* de Pearson
RS		*r* de Spearman
TAU		*tau* de Kendall
KHI		Statistique *G* (khi carré de Wilks)
HT		Coefficient de contingence de Tschuproff
HS0	B/(A+B+C)	Coefficient d'information réciproque (Estabrook)
HS1	√[1 - (HD)**2]	Coefficient de cohérence (Rajski)
HS2	B/(A+2B+C)	Coefficient symétrique d'incertitude
HD	(A+C)/(A+B+C)	Métrique de Rajski

Tableau 5 - Le choix d'une mesure d'association asymétrique entre objets (mode Q) pour tableau d'abondances d'espèces ou autres descripteurs pour lesquels les doubles zéros ne sont pas indicateurs de ressemblance. Modifié de Legendre & Legendre (1984a), tableau 7.III.

1) Données de présence-absence, ou échelle d'abondance relative sans similarité partielle entre les classes	voir 2
2) Coefficients métriques: S07, S10, S11, S26	
2) Coefficients semi-métriques: S08, S09, S13, S14	
2) Coefficient non-métrique: S12	
1) Données quantitatives	voir 3
3) Données brutes	voir 4
4) Sans niveau de probabilité	voir 5
5) Sans standardisation par objet; une même différence entre deux objets, pour des espèces abondantes ou rares, a la même contribution à la similarité: S17, S18	
5) Standardisation par vecteur-objet; les différences entre objets pour les espèces les plus abondantes (dans l'ensemble du fichier) contribuent davantage à la similarité (moins à la distance): S21	
4) Coefficient probabiliste: S22	
3) Données normalisées (ou, du moins, distribution non asymétrique) ou sur échelle d'abondance relative	voir 6
6) Sans niveau de probabilité	voir 7
7) Sans standardisation par objet	voir 8
8) Une même différence entre les deux objets, pour des espèces abondantes ou rares, a la même contribution à la similarité: S17, S18, D08, D14	
8) Les différences entre objets pour les espèces abondantes (dans les deux objets considérés) contribuent davantage à la similarité (moins à la distance): D10, D11	
8) Les différences entre objets pour les espèces les plus abondantes (dans l'ensemble du fichier) contribuent davantage à la similarité (moins à la distance): S19, S20	
7) Standardisation par vecteur-objet; pour des objets d'importance égale, ces mesures donnent la même contribution aux espèces abondantes ou rares: D03, D04 (où l'importance se calcule par la longueur du vecteur), D09 (où elle se calcule par l'effectif total du vecteur)	
6) Coefficient probabiliste: S23	

Fichiers d'entrée et de sortie

(1) Fichier d'entrée principal

Dans le fichier d'entrée, les données sont des nombres entiers ou réels, positifs ou négatifs, écrits en ASCII. Le programme SIMIL calcule toujours ses mesures de ressemblance entre les lignes du fichier d'entrée; il faudra s'assurer que les lignes représentent les objets si on désire calculer une similarité ou une distance (mode Q d'analyse), et les descripteurs si on désire calculer un coefficient de dépendance (mode R). Le programme VERNORM permet de vérifier le contenu des fichiers de données, de transposer les matrices et de normaliser les descripteurs, si besoin est:

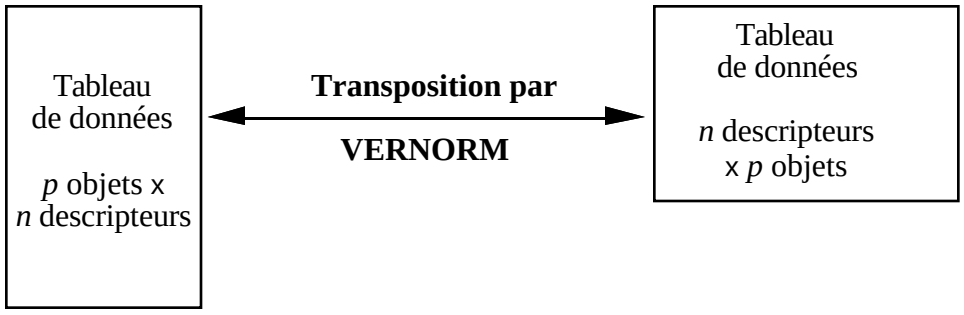


Tableau 6 - Le choix d'une mesure d'association symétrique entre objets (mode Q) pour tableau de descripteurs physiques, chimiques, géologiques, etc. Modifié de Legendre & Legendre (1984a), tableau 7.IV.

1) Comparaison d'objets individuels	voir 2
2) Descripteurs binaires, ou à descriptions multiples sans similarités partielles	voir 3
3) Coefficients métriques: S01, S02, S06	
3) Coefficients semi-métriques: S03, S05	
3) Coefficient non-métrique: S04	
2) Descripteurs à descriptions multiples	voir 4
4) Descripteurs quantitatifs, dimensionnellement homogènes	voir 5
5) Différences soulignées par mise au carré: D01, D02	
5) Différence atténuée: D07, D08	
4) Descripteurs sans homogénéité dimensionnelle; des poids égaux (ou différents, selon les valeurs w_i imposées) sont attribués aux différents descripteurs	voir 6
6) Descripteurs qualitatifs (sans similarités partielles) et descripteurs quantitatifs avec similarités partielles basées sur l'écart de variation de chaque descripteur: S15	
6) Descripteurs qualitatifs (possibilité de matrices de similarités partielles entre les classes) et descripteurs quantitatifs ou semi-quantitatifs avec fonction de similarité partielle pour chaque descripteur: S16	
1) Comparaison de groupes d'objets	voir 7
7) Tenant compte de la corrélation entre descripteurs: D05	
7) Sans tenir compte de la corrélation entre descripteurs: D12	

Pour le mode Q d'analyse, les objets forment les lignes successivement de cette matrice; sur une ligne, les différents descripteurs sont inscrits en ordre, à la suite. Cependant, un objet peut prendre autant de lignes dans le fichier qu'il est nécessaire pour accommoder tous ses descripteurs. Comme la lecture des données sera effectuée en format libre, les descripteurs doivent être séparés par un ou plusieurs espaces (le nombre d'espaces n'a pas d'importance; à la limite, on peut même n'inscrire qu'une donnée par ligne du fichier). Une conséquence de cette flexibilité de lecture est que les absences d'informations ne peuvent être représentées par des espaces blancs, qui sont ignorés lors de la lecture des données; les informations absentes doivent être matérialisées par un code numérique (on emploie souvent 0, -1, -9 ou -999), qui sera déclaré en réponse à une question du programme. Ce code doit différer de façon non-ambiguë de toute valeur numérique pouvant légitimement se trouver dans le fichier.

Dans le cas des coefficients de distances D05 et D12, le calcul se fera entre des groupes d'objets. Il est nécessaire que les objets membres d'un même groupe se retrouvent l'un à la suite de l'autre dans le fichier d'entrée. On ne peut spécifier à l'aide d'un code à quel groupe appartient chaque objet; le programme demandera de donner, dans l'ordre, le nombre d'objets membres de chaque groupe.

Il n'y a en principe pas de limite quant à la taille des matrices qui peuvent être traitées par la version Macintosh de ce programme. Le programme occupe tout l'espace mémoire (RAM) qui lui est disponible, si bien que la taille des matrices que le programme peut traiter en pratique sera une fonction, non seulement de la taille de la mémoire disponible dans la machine, mais également de la version du Système utilisée ainsi que de l'utilisation simultanée de MultiFinder, d'une mémoire-cache ou d'autres programmes. Les versions 3.0 et plus de SIMIL réalisent tous les calculs en mémoire centrale afin de les accélérer; s'il n'y a pas suffisamment d'espace disponible pour traiter le tableau de données, le message suivant sera émis:

```
Manque de mémoire! Essayez une ancienne version de SIMIL
```

Tableau 7 - Le choix d'un coefficient de dépendance entre descripteurs (mode R). Modifié de Legendre & Legendre (1984a), tableau 7.V.

1) Descripteurs: abondances d'espèces	voir 2
2) Données brutes: S21, RS, TAU	
2) Données normalisées	voir 3
3) Sans niveau de probabilité: RP (après avoir éliminé des doubles zéros, autant que possible); RS, TAU	
3) Coefficients probabilistes: probabilité associée à RP, RS et TAU; S23	
2) Données de présence-absence	voir 4
4) Sans niveau de probabilité: S7, S8, S24	
4) Coefficient probabiliste: S25	
1) Autres descripteurs: physiques, chimiques, géologiques, etc.	voir 5
5) Sans niveau de probabilité	voir 6
6) Descripteurs quantitatifs en relation linéaire: RP	
6) Autres descripteurs ordonnés, en relation monotone: RS, TAU	
6) Descripteurs ordonnés en relation non monotone et descripteurs qualitatifs: KHI, HT, HS0, HS1, HS2, HD	
5) Coefficients probabilistes	voir 7
7) Descripteurs quantitatifs en relation linéaire: probabilité associée à RP	
7) Autres descripteurs ordonnés, en relation monotone: probabilité associée à RS, TAU	
7) Descripteurs ordonnés en relation non monotone et descripteurs qualitatifs: probabilité associée à KHI	

La première solution consiste à quitter MultiFinder si on s'y trouve; si le problème ne s'en trouve pas résolu, on peut tenter d'employer une version de SIMIL d'un numéro inférieur à 3; celles-ci gardent sur disque la plus grande partie du tableau de données et peuvent donc traiter des tableaux plus grands, au prix d'une rapidité d'exécution moins grande. Notez que les versions de SIMIL d'un numéro inférieur à 3 sont gourmandes en espace-disque nécessaire à l'exécution du programme, ce qui a justifié la mise au point de la version 3. Quant aux versions CMS et VMS, la taille maximale des fichiers qui peuvent être traités est limitée par les paramètres inscrits au début du programme; l'usager pourra les ajuster à ses besoins avant la compilation.

Le programme permet à l'usager qui le désire d'inscrire un identificateur au début de chaque vecteur-objet, mais pas en tête des colonnes. Si on déclare au programme qu'il y a de tels identificateurs, le programme assumera que ceux-ci occupent les 10 premiers caractères de chaque vecteur-objet (ligne ou ensemble de lignes); tout caractère alphanumérique peut être employé pour ces identificateurs, y compris les blancs. Dans ce cas, la liste des descripteurs commencera en colonne 11 ou plus loin. Cette convention est la même que celle du progiciel d'analyse phylogénétique PHYLIP du Prof. Joseph Felsenstein. Le fichier suivant, avec identificateurs, serait un fichier acceptable pour SIMIL (résultats de pêche: 6 objets, 4 descripteurs; l'absence d'information est notée -9); les 10 espaces réservés aux identificateurs sont matérialisés ci-dessous par un souligné:

```
poisson1  1 3.2 4 5
poisson2    1          2.9           3     4
b.conserve
2 0.9 -9 -9
sac Glad
2
15.0
-9 -9
poisson3
          1 3.5 4 20
vieux pneu2 75.4 -9 -9
```

Notons cependant que les fichiers de données soumis à SIMIL sont souvent extraits de fichiers plus grands gérés par des programmes de bases de données ou des logiciels statistiques; ils ont donc plus souvent l'apparence suivante, s'ils présentent des identificateurs dans les colonnes 1 à 10:

```
Stat.100    2    4    3    1    1    4    1    1    1    2    2
Stat.200    2    4    3    1    1    4    1    1    1    2    2
Stat.320    2    4    3    1    1    4    1    1    1    1    4
Stat.330    2    4    3    1    1    4    1    1    1    1    4
Stat.340    2    4    3    1    1    4    1    1    1    1    4
```

ou encore, sans identificateurs:

```
-0.38566  -1.42712  37.1  8.24931  0.02627  0.85015
-0.01005   0.77932  37.5  7.34987  0.01033  0.77932
 0.10436   0.94391  37.1  7.09589  0.16279  0.49348
 0.33647   0.71295  37.4  6.79571  0.09373  0.57098
 0.30748   0.52473  37.3  6.57508  0.14691  1.39128
```

ATTENTION: Comme pour tous les autres programmes de ce progiciel, il faut écrire par exemple "0.376" et non pas ".376", et "-0.42" et non pas "-.42" lorsqu'on utilise une version de SIMIL antérieure à la version 3, ou encore les versions CMS ou VMS; voir la remarque à cet effet à la page 6.

(2) Fichier d'entrée des matrices de similarités partielles

Pour les coefficients S16 et S20, le programme demande combien on a de matrices de similarités partielles. Si de telles matrices sont utilisées pour quantifier les relations entre les classes de variables semi-quantitatives, qualitatives ou circulaires (Legendre & Legendre, 1984a), il faut créer un second fichier qui, pour chaque matrice de similarités partielles, doit contenir les informations suivantes:

1- Le numéro du descripteur (colonne) auquel s'applique cette matrice.
2- La taille de la matrice partielle, qui est égale au nombre maximum de valeurs (ou classes) que peut prendre ce descripteur.
3- Cette matrice elle-même, en nombres réels, sous forme triangulaire inférieure, diagonale exclue. Si le descripteur qualitatif en question comporte *n* classes, il doit donc y avoir (*n**(*n*-1))/2 valeurs dans la matrice de similarités partielles.

Supposons par exemple que les descripteurs 2 et 4 demandent les matrices suivantes de similarités partielles (chacune étant ici d'ordre 5):

descripteur 2 (5 classes):

```
  1
0.4    1
0.5   0.6    1
0.5   0.4   0.45   1
0.46  0.47  0.5    0.5    1
```

descripteur 4 (5 classes):

```
  1
0.4    1
0.3   0.8    1
0.9   0.2   0.55   1
0.48  0.9   0.2    0.8    1
```

Le fichier d'entrée des matrices de similarités partielles serait alors le suivant:

```
2  5
0.4
0.5  0.6
0.5  0.4  0.45
0.46  0.47  0.5  0.5
4  5
0.4  0.3  0.8  0.9  0.2  0.55  0.48  0.9  0.2  0.8
```

La première ligne donne le numéro de descripteur et le nombre de classes pour la première matrice partielle; la ligne 6 donne les mêmes informations pour la seconde. Les lignes successives de la première matrice sont écrites sur des lignes physiques distinctes, alors que les lignes successives de la seconde matrice sont écrites à la suite, sur une seule ligne physique; les deux formats sont acceptables.

(3) Fichier de sortie principal

Le fichier de sortie contenant les résultats des calculs, décrit ci-dessous, est écrit en binaire, prêt à être relu par les différents programmes d'analyse de données du progiciel *R* (voir pages 3 et 7 de ce manuel). Il est possible de lire le contenu de cette matrice binaire grâce au programme utilitaire REGARDE qui la transcrira en caractères lisibles, ou encore à l'aide de la procédure ci-dessous. Notez que les matrices binaires construites par SIMIL sur un type d'ordinateur ne peuvent pas être utilisées par les programmes de *R* résidant sur un autre type d'ordinateur, à cause des différences qui existent entre les machines des différentes marques quant à la structure des mots-machine.

La structure de ce fichier binaire est la suivante:

- Premier mot (entier): nombre de lignes (ou de blocs de lignes) entre lesquels les mesures de ressemblance ont été calculées.

- Second mot (entier): Nombre de colonnes dans le fichier des données brutes.

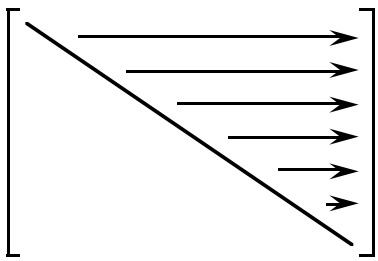
- Mots 3 à 12 (caractères): Titre donné au fichier par l'usager.

- Mot 13 (caractères): Date de création du fichier.

- Mots 14 et 15 (caractères): Nom de la mesure de ressemblance employée.

- Mot 16 (entier): Nombre k d'identificateurs à lire; s'il n'y avait pas d'identificateurs au début des lignes du tableau des données, $k = 0$. Si $k > 0$, cette information est suivie d'une liste de k mots-machine contenant chacun un identificateur de ligne (10 caractères).

- Suit la liste des ressemblances, selon les lignes successives de la matrice triangulaire supérieure, à l'exclusion de la diagonale. Un mot-machine (nombre réel) est utilisé pour chaque mesure de ressemblance.



La procédure PASCAL qui suit indique comment lire cette matrice binaire de similarités. Notez que cette procédure est écrite pour CMS; si on veut l'employer pour relire les matrices de type SIMIL produites en version Macintosh, il faut savoir qu'il faut écrire `"READ(ENTREE,VAL);"` par exemple, plutôt que `"VAL:=ENTREE@;GET(ENTREE);"`. De plus, alors que les versions CMS et VMS de SIMIL utilisent des réels de 8 bytes, la version Macintosh de SIMIL utilise des réels de 10 bytes (définis comme "Extended" en MPW PASCAL). L'usager suffisamment familier avec PASCAL pour vouloir incorporer cette procédure dans son propre programme connaîtra vraisemblablement les autres adaptations qui sont nécessaires pour sa machine ainsi que pour le compilateur qu'il utilise.

```
PROCEDURE TypeSIMIL;
CONST NBMOTS=1;
      NBCAR=10;
(* ================================================================ *)
(* Cette procédure lit une matrice binaire de type SIMIL et la      *)
(* transcrit en clair. Les valeurs de ressemblance sont également   *)
(* écrites en binaire dans une matrice carrée MAT.                  *)
(* ================================================================ *)

TYPE MATRICE = ARRAY[1..500, 1..500] OF REAL;
     VARIABLE= RECORD CASE INTEGER OF
               1:(INT:INTEGER);
               2:(RE:REAL);
               3:(CAR:PACKED ARRAY[1..NBCAR] OF CHAR);
               END;
VAR  I,J,NOBJ,NBDESC:INTEGER;
     ENTREE: FILE OF VARIABLE;
     MAT: MATRICE;
     VAL: VARIABLE;

   PROCEDURE ECRITDATA;
   VAR I,J,K:INTEGER;
   BEGIN
   J:=0;
   FOR I:=1 TO NBMOTS DO
      BEGIN
      VAL:=ENTREE@;GET(ENTREE);
      FOR K:=1 TO NBCAR DO
            BEGIN
            J:=J+1;
            IF J<=10 THEN WRITE(SORTIE,VAL.CAR [K]);
            END;
      END;
   END;                                                  (* Fin de ECRITDATA *)

BEGIN
RESET(ENTREE);
WRITELN(SORTIE,'FICHIER D''ENTREE:');
VAL:=ENTREE@;GET(ENTREE);
NOBJ:=VAL.INT;                  (* NOBJ: nombre de lignes de la matrice de données *)
WRITELN(SORTIE,'    NOMBRE D''OBJETS: ',NOBJ:4);       (* Impression de NOBJ *)
VAL:=ENTREE@;GET(ENTREE);
NBVAR:=VAL.INT;                                  (* NBVAR: nombre de colonnes *)
WRITELN(SORTIE,'    NOMBRE DE VARIABLES: ',NBVAR:4); (* Impression de NBVAR *)
WRITE  (SORTIE,'    TITRE: ');
```

```
FOR I:=1 TO 10 DO ECRITDATA;                                    (* Impression du titre *)
WRITELN(SORTIE);
WRITE  (SORTIE,'    DATE: ');
ECRITDATA;                                                      (* Impression de la date *)
WRITELN(SORTIE);
WRITE  (SORTIE,'    MESURE DE RESSEMBLANCE: ');
ECRITDATA;ECRITDATA;                        (* Impression de la mesure de ressemblance *)
WRITELN(SORTIE);
WRITELN(SORTIE); WRITELN(SORTIE,'LISTE DES IDENTIFICATEURS:');
VAL:=ENTREE@;GET(ENTREE);                 (* Lecture du nombre d'identificateurs présents *)
IF VAL.INT<> 0 THEN                                    (* Lecture des identificateurs *)
      BEGIN
      FOR I:=1 TO NOBJ DO
            BEGIN
            FOR J:=1 TO NBMOTS DO ECRITDATA;      (* Impression des identificateurs *)
            WRITELN(SORTIE);
            END;
      END;
WRITELN(SORTIE); WRITELN(SORTIE,'LISTE DES VALEURS DE RESSEMBLANCE:');
FOR I:=1 TO NOBJ DO                           (*  Lecture des valeurs de ressemblance *)
      BEGIN
      MAT[I,I]:=1.0;
      FOR J:=I+1 TO NOBJ DO
            BEGIN
            VAL:=ENTREE@;GET(ENTREE);
            WRITELN(SORTIE,VAL.RE:10:5);(* Impression des valeurs de ressemblance *)
            MAT[I,J]:=VAL.RE;               (* Valeurs écrites en binaire dans MAT *)
            MAT[J,I]:=MAT[I,J];
            END;
      END;
END;
```

(4) Fichier de sortie de la matrice de probabilités

Lorsqu'un test de signification statistique est associé à un coefficient (S23, S25, D05, RP, RS, TAU, KHI), une matrice des probabilités écrite également en binaire peut être obtenue dans un second fichier de sortie, à la demande de l'usager. Par défaut, ce fichier de sortie est appelé "PROBAB DATA A" dans la version CMS. C'est le complément de la probabilité de l'hypothèse nulle qui est notée, en fait, dans ce fichier **[revoir ?]**. Ainsi, on obtient des valeurs élevées pour un coefficient significatif et une valeur basse lorsqu'il n'y a pas de relation entre les deux objets ou descripteurs en question; ces valeurs de probabilités se comportent donc comme des mesures de similarité et peuvent être utilisées comme tel dans les programmes de groupement ou d'ordination. Contrairement aux coefficients de corrélation, par exemple, où les deux extrêmes de l'échelle correspondent à l'existence d'une relation entre deux descripteurs, les valeurs de probabilité ont, quant à elles, un comportement monotone sur leur échelle de variation [0, 1]. On peut consulter ce fichier à l'aide du programme REGARDE.

Les options du programme

Le programme offre comme options les 50 mesures de ressemblance énumérées au tableau 4. L'exposé détaillé de ces mesures dépasse le cadre du présent document; on pourra se référer au texte de Legendre & Legendre (1984a) ou à l'une des revues suivantes de certains coefficients de ressemblance: Sokal & Sneath (1963), Williams & Dale (1965), Cheetham & Hazel (1969), Sneath & Sokal (1973), Clifford & Stephenson (1975), Daget (1976), Blanc *et al*. (1976), Orlóci (1978),

Gower (1985). Les critères devant guider l'usager dans le choix d'un coefficient sont résumés aux tableaux 5 à 7. La section suivante montre en quoi le dialogue du programme avec l'usager diffère selon la mesure de ressemblance sélectionnée.

Les questions du programme

Les questions présentées par le programme à l'écran du Macintosh sont décrites dans les paragraphes qui suivent. Les questions posées par les versions CMS et VMS sont essentiellement les mêmes, comme on peut le vérifier avec l'exemple présenté à la section suivante. Pour faire démarrer le programme, il faut cliquer sur l'icône, puis donner la commande "Ouvrir" dans le menu "Fichiers".

(1) "Fichier d'entrée" — Le programme présente la liste des fichiers ASCII disponibles.

(2) "Titre:" — On fournit un titre, qui sera inscrit dans le bloc d'informations de chaque fichier de ressemblance produit par SIMIL; voir le programme REGARDE pour détails.

(3) "Nombre de lignes (ou de blocs de lignes)" — La réponse doit être un nombre entier positif. En mode Q, il s'agit du nombre d'objets, chaque objet pouvant occuper une ou plusieurs lignes; en mode R, où la matrice est transposée, il s'agit du nombre de variables, chaque variable pouvant également occuper une ou plusieurs lignes; voir la description du fichier d'entrée principal.

(4) "Nombre de colonnes" — En mode Q, on inscrit le nombre de descripteurs décrivant chaque objet du fichier, à l'exclusion des identificateurs de lignes si le fichier en possède. En mode R, où le fichier est transposé, on inscrit le nombre d'objets composant chaque vecteur-variable, à l'exclusion des identificateurs de descripteurs.

(5) "Code indiquant l'absence d'information (par défaut: 0)" — On inscrit quelle valeur numérique a été utilisée dans le fichier pour indiquer qu'une information est absente (souvent: -1, -9, -999, etc.). On doit répondre à cette question par une valeur numérique même s'il n'y a pas de données manquantes dans le fichier.

(6) "Les 10 premiers caractères de chaque ligne sont des identificateurs" [Oui, Non] — On répond *Oui* si les 10 premières colonnes de chaque vecteur-objet ou descripteur contient un identificateur de cet objet ou descripteur; voir la section sur le fichier d'entrée principal.

(7) "Calculs" [Similarités, Distances, Autres; Information] — Si on choisit les Similarités, un nouveau menu offre le choix entre les similarités S1 à S26 du tableau 4; si on choisit Distances, le menu présenté donne le choix entre les distances D1 à D14; enfin, Autres mène à un nouveau menu offrant le choix entre les fonctions Tau, R de Pearson, R de Spearman, Khi, Ht, Hs0, Hs1, Hs2, Hd et Nei. Informations donne accès à un fichier contenant le tableau 4 lui-même. On peut monter ou descendre dans ce tableau en pointant le curseur de la souris dans le haut ou le bas de l'écran; le tableau peut également être envoyé à l'imprimante si l'usager le désire.

Dans les versions CMS et VMS, la question se présente de la façon suivante, qui contient les mêmes possibilités de choix:

```
QUELLE FONCTION DESIREZ-VOUS CALCULER ?
Similarites   : s01  a  s26, ou  Nei
Distances     : d01  a  d14
Mode R:    rp = r de Pearson
           rs = r de Spearman
           tau= Tau de Kendall
           khi= Khi-carre (statistique G)
```

```
ht = Coefficient de contingence de Tschuproff
hs0= Information reciproque         S=B/(A+B+C)
hs1= Coherence de Rajski     S'=SQRT(1-(hd)**2)
hs2= Coeff. symetr. d'incertitude S"=B/(A+2B+C)
hd = Metrique de Rajski        D =(A+C)/(A+B+C)
```

(8) “Fichier de sortie” — Le programme présente le menu permettant de donner un nom au fichier binaire de mesures de ressemblance qu’il produira.

À partir de ce point, les questions divergent selon le type de coefficient que l’on désire calculer. Ces questions sont suivies du calcul des coefficients, après quoi on retourne au menu “Fichiers” qui permet de traiter immédiatement un autre fichier de données. La commande “Interrompre” dans le menu “R: Simil” permet de quitter le programme.

Les coefficients binaires: S1 à S14, S24, S25, S26, D13

(9) “Seuil à partir duquel l’information sera codée 1 (plus petit: 0)” — Si le fichier de données ne contient que des “0” et des “1”, on répond “1” à cette question. Si par ailleurs le fichier contient des données quantitatives, il est possible de les faire traiter par le programme comme s’il s’agissait de données de présence-absence. On aurait pu établir que toute valeur plus grande que zéro doit être recodée “1”; avec des données d’abondance d’espèces par exemple, il se peut cependant que l’usager décide de considérer comme absente toute espèce qui n’est pas représentée, par exemple, par au moins 10 individus à la station d’échantillonnage; sa réponse serait alors “10”. En général, la réponse à cette question indique à partir de quelle valeur numérique l’usager demande au programme de considérer l’espèce comme “présente”. La plus petite valeur admissible est “0”.

Les coefficients quantitatifs simples: D1 à D4, D7 à D11, D14, NEI

Aucune question supplémentaire n’est posée par le programme avant le calcul de ces coefficients.

La métrique de Minkowski: D6

(9) “Puissance pour cette fonction” — On répond par un entier positif, qui fournit l’exposant “r” de la métrique de Minkowski dont la formule est $D6(\mathbf{x}_1, \mathbf{x}_2) = [\ \Sigma\ |y_{i1} - y_{i2}|^r\ ]^{(1/r)}$. La métrique de Manhattan (D7) correspond à la métrique de Minkowski à l’exposant 1, alors que la distance euclidienne (D1) est la métrique de Minkowski à l’exposant 2.

Les distances entre groupes d’objets: D5 et D12

(9) “Fichier de probabilités associées” [Oui, Non] — Cette question n’est posée que pour le coefficient D5 (distance généralisée de Mahalanobis), car le coefficient D12 (coefficient de ressemblance raciale) ne conduit pas à un test de signification statistique des différences calculées entre les groupes. Si on répond *Oui*, le programme demande le nom que doit recevoir le fichier binaire de probabilités.

(10) “Cardinalité du groupe 1” — On indique ici le nombre (entier positif) d’objets dans le premier groupe. Le programme demande ensuite la “Cardinalité du groupe 2”, etc. jusqu’à épuisement des objets dont le nombre a été fourni en réponse à la question 3.

Les coefficients de Gower: S15 et S19

(9) “Ecart sur les données plutôt que celui de la population” [Oui, Non] — Le coefficient de Gower a comme formule $D(\mathbf{x}_1, \mathbf{x}_2) = \Sigma\ w_{i12}\, s_{i12} / \Sigma w_{i12}$. Les poids w_i seront traités à la question (10). Nous nous intéressons ici à la fonction de similarité partielle s_{i12} entre les objets $\mathbf{x}_1$ et $\mathbf{x}_2$ pour les descripteurs quantitatifs. Dans ce cas, la différence entre les valeurs du descripteur pour ces deux

objets, $|y_{i1} - y_{i2}|$, est rapportée à l'écart maximum R_i que peuvent prendre les valeurs de ce descripteur; la question cherche à déterminer si cet écart R_i doit être calculé à partir du tableau de données lui-même (réponse *Oui*), ou si l'usager désire fournir lui-même les valeurs d'écart R_i qu'il connaît par ailleurs à partir de la population de référence dont est extrait l'échantillon à l'étude (réponse *Non*). La similarité partielle s_{i12} entre les objets $\mathbf{x}_1$ et $\mathbf{x}_2$ est calculée par $s_{i12} = 1 - [\,|y_{i1} - y_{i2}| / R_i\,]$.

(9.1) "Ecart pour variable 1" — Si on a répondu *Non* à la question précédente, indiquant par là que l'on désire fournir les valeurs de R_i, le programme demande maintenant la valeur de l'écart pour la première variable (nombre réel positif). Le programme demande ensuite: "Ecart pour variable 2", etc. jusqu'à épuisement des variables. On doit fournir une valeur-bidon pour les variables qualitatives multiclasses, qui ne seront identifiées qu'à la question (11).

(10) "Tous les poids (W[i]) sont binaires (0 ou 1)" [Oui, Non] — Les valeurs w_i ont deux rôles distincts dans la formule de ces coefficients. D'une part, elles servent à donner des poids variables aux différents descripteurs, si l'usager le désire; si on ne désire pas se prévaloir de cette option, on répond *Oui* à la question, ce qui donne par défaut des poids égaux à tous les descripteurs dans le calcul de la similarité globale. Le second rôle de ces valeurs est de permettre d'éliminer du calcul global tout descripteur pour lequel l'un des deux objets souffre d'une absence d'information (dont le code a été établi en réponse à la question 5); dès qu'il y a absence d'information, le descripteur se voit attribuer un poids $w_i = 0$. Enfin, dans la forme asymétrique du coefficient (S19), $w_i = 0$ lorsque l'espèce est absente des deux vecteurs-objets ($y_{i1} + y_{i2} = 0$).

(10.1) "W[1]" — Si on a répondu *Non* à la question (10), on doit maintenant indiquer le poids désiré pour le descripteur no 1. La réponse doit être un nombre réel ≥ 0; un poids de zéro équivaut à éliminer le descripteur des calculs. Le programme demande ensuite: "W[2]", etc. jusqu'à épuisement des variables.

(11) "Nombre de descripteurs qualitatifs multiclasses" — Les descripteurs qualitatifs multiclasses sont traités par le coefficient S15 de la même façon que le coefficient de simple concordance pour données multiclasses: on compte une similarité partielle $s_{i12} = 1$ s'il y a accord entre les deux objets pour ce descripteur, et 0 s'il y a désaccord. Si le fichier contient des descripteurs qualitatifs qui doivent être traités de cette façon, on doit indiquer ici combien il y a de descripteurs de ce type. Cette question n'est posée que pour la forme symétrique du coefficient (S15); dans la forme asymétrique (S19), réservée aux données de fréquence (abondances d'espèces, en écologie), cette question n'aurait pas de sens.

(11.1) "Identificateur du descripteur 1" — On indique ici quel est le numéro du premier descripteur qualitatif, parmi les descripteurs du fichier de données. Le programme demande ensuite: "Identificateur du descripteur 2", etc. jusqu'à épuisement des descripteurs qualitatifs. Si le nombre déclaré de descripteurs qualitatifs (question 11) est égal au nombre total de descripteurs (question 4), cette question n'est pas posée.

Les coefficients S16 et S20

(9) "Nombre de matrices de similarités partielles" — Les coefficients S16 et S20 ont la même formule générale que les coefficients de Gower, soit $D(\mathbf{x}_1, \mathbf{x}_2) = \Sigma\, w_{i12}\, s'_{i12} / \Sigma w_{i12}$. Les poids w_i seront traités à la question (11). Nous nous intéressons ici à la fonction de similarité partielle s'_{i12} entre les objets $\mathbf{x}_1$ et $\mathbf{x}_2$; c'est là que S16 et S20 diffèrent de S15 et S19 respectivement. La valeur de s'_i peut être déterminée de deux façons différentes: soit par une fonction monotone décroissante décrite à la question (10), soit en imposant des valeurs de similarités partielles entre les différentes classes d'un descripteur qualitatif, semi-quantitatif ou circulaire. Legendre & Legendre (1984a, tome 2, p. 15) fournissent un exemple d'une telle matrice de similarités partielles. Ces matrices, une pour chacun des descripteurs devant être traité de cette façon, doivent être inscrites l'une à la suite de l'autre dans un fichier séparé, en suivant les indications fournies à la section "Fichier d'entrée des matrices de similarités partielles". La réponse doit être un entier ≥ 0; on répond "0" (zéro) si aucune matrice de

similarités partielles n'est fournie. Si on répond par un entier positif, le programme présente un menu des fichiers ASCII disponibles; on indiquera lequel contient les matrices de similarités partielles. Ces matrices représentent le seul moyen disponible d'imposer des similarités partielles entre les classes d'un descripteur qualitatif ou d'une variable circulaire. Les coefficient S16 et S20 sont très utiles pour le traitement des tableaux comportant des descripteurs appartenant à plusieurs types (quantitatifs, semi-quantitatifs, qualitatifs).

(10) "Même valeur de K[i] pour tous les descripteurs" [Oui, Non] — Estabrook & Rogers (1966) ont proposé d'estimer la similarité partielle entre les valeurs d'un descripteur quantitatif à l'aide d'une fonction de similarité partielle empirique qui est une fonction à la fois de la distance $d_{i12} = |y_{i1} - y_{i2}|$ entre les valeurs prises par deux objets pour ce descripteur et d'une borne k_i fixée par l'usager, borne qui limite l'extension de la similarité partielle à une distance maximum k_i. Cette fonction empirique a pour équation $s'_{i12} = f(d_{i12}, k_i) = 2 (k_i + 1 - d_{i12}) / (2 k_i + 2 + d_{i12} k_i)$ si $d_{i12} \leq k_i$ et $s'_{i12} = 0$ lorsque $d_{i12} > k_i$. De plus, avec des données d'abondance d'espèces (coefficient S20), $s'_{i12} = 0$ lorsque y_{i1} ou y_{i2} ont la valeur zéro (Legendre & Chodorowski, 1977). Des exemples d'utilisation de cette fonction se trouvent dans l'article de Estabrook & Rogers (1966) ainsi que dans Legendre & Legendre (1984a, tome 2, p. 13-14). Cette question du programme cherche à déterminer si des valeurs différentes de k_i seront attribuées aux différents descripteurs.

(10.1) "K[i]" — Si on a répondu *Oui* à la question (10), on fournit ici la valeur unique de k_i qui sera utilisée pour tous les descripteurs. Cette valeur est un nombre réel ≥ 0. Lorsque $k_i = 0$, le descripteur est traité de la même façon que le coefficient de simple concordance (dans le cas de S16) ou le coefficient de Jaccard (dans le cas de S20) pour données multiclasses: on compte une similarité partielle $s_{i12} = 1$ s'il y a accord entre les deux objets pour ce descripteur, et 0 s'il y a désaccord.

(10.2) "K[1]" — Si on a répondu *Non* à la question (10), on fournit ici la valeur de k qui sera utilisée pour le premier descripteur. Le programme demande ensuite: "K[2]", etc. jusqu'à épuisement des descripteurs. On doit fournir une valeur-bidon pour les variables qualitatives multiclasses, qui faisaient l'objet de la question (9).

(11) "Tous les poids (W[i]) sont binaires (0 ou 1)" [Oui, Non] — Les valeurs w_i ont deux rôles distincts dans la formule de ces coefficients. D'une part, elles servent à donner des poids variables aux différents descripteurs, si l'usager le désire; si on ne désire pas se prévaloir de cette option, on répond *Oui* à la question, ce qui donne par défaut des poids égaux à tous les descripteurs dans le calcul de la similarité globale. Le second rôle de ces valeurs est de permettre d'éliminer du calcul global tout descripteur pour lequel l'un des deux objets souffre d'une absence d'information (dont le code a été établi en réponse à la question 5); dès qu'il y a absence d'information, le descripteur se voit attribuer un poids $w_i = 0$. Enfin, dans la forme asymétrique du coefficient (S20), $w_i = 0$ lorsque l'espèce est absente des deux vecteurs-objets ($y_{i1} + y_{i2} = 0$).

(11.1) "W[1]" — Si on a répondu *Non* à la question (11), on doit maintenant indiquer le poids désiré pour le descripteur no 1. La réponse doit être un nombre réel ≥ 0; un poids de zéro équivaut à éliminer le descripteur des calculs. Le programme demande ensuite: "W[2]", etc. jusqu'à épuisement des variables.

La similarité probabiliste du khi carré: S22

(9) "Khi carré de Wilks plutôt que de Pearson" [Oui, Non] — Le coefficient S22 est le complément de la probabilité associée à la statistique khi carré calculée sur le tableau de fréquences formé par deux échantillons et *n* espèces, après avoir exclu les doubles zéros des calculs. L'usager a le choix entre la statistique khi carré de Wilks (ou statistique *G*: répondre *Oui*) ou la statistique de Pearson (répondre *Non*).

Le coefficient probabiliste de Goodall: S23

(9) “Calcul sur indice de Gower plutôt que Steinhaus” [Oui, Non] — Le coefficient S23 est le complément de la probabilité que deux échantillons pris au hasard soient aussi similaires ou plus similaires que la paire d’échantillons en question. Les similarités partielles par espèce sur lesquelles sont basés les calculs peuvent être calculées à la façon de l’indice de Gower S19 (réponse *Oui*) ou à celle de l’indice de Steinhaus S17 (réponse *Non*), tel qu’expliqué par Legendre & Legendre (1984a, tome 2, p. 22).

Les coefficients de dépendance entre descripteurs (mode R): de RP à HD

(9) “Fichier de probabilités associées” [Oui, Non] — Si on désire obtenir le fichier des probabilités associées, on répond *Oui* à cette question; dans ce cas, le programme demande de donner un nom à ce fichier binaire, qui est décrit à la section “Fichier de sortie de la matrice de probabilités”. On peut consulter ce fichier à l’aide du programme REGARDE.

(10) “Calcul de Tau A et B plutôt que Tau A, B & C” [Oui, Non] — Il existe trois versions du coefficient de corrélation non-paramétrique *tau* de Kendall: tau_a est employé lorsqu’il n’y a pas d’observations liées, tau_b lorsqu’il y a des observations liées et que les deux variables comportent le même nombre de classes semi-quantitatives, et tau_c lorsqu’il y a des observations liées mais que le nombre de classes n’est pas le même pour les deux descripteurs. Le programme SIMIL choisit la version qui convient dans chaque situation. Cependant, certains auteurs recommandent de ne plus employer la formule de correction du tau_c, mais d’utiliser plutôt tau_b dans tous les cas où il y a des observations liées; l’usager de SIMIL peut ici décider de ne faire calculer que tau_a et tau_b (réponse *Oui*) s’il le désire. Cette question n’est posée que si le coefficient choisi est le *tau* de Kendall.

Exemple

Voici un exemple d’utilisation du programme SIMIL sur grands ordinateurs. Le fichier de données contient 71 objets et 11 descripteurs; l’absence d’information est codée “-9”. Le coefficient S15 sera calculé entre les lignes de ce fichier; voir les questions particulières à cette fonction, à la section précédente. Ce coefficient ne permet pas d’avoir recours à des matrices de similarités partielles; donc, aucune réponse n’est fournie lorsqu’un nom est demandé pour ce fichier (voir 1 en marge gauche). De même, aucune réponse ne sera fournie à la demande d’un nom pour le fichier des probabilités puisque le coefficient S15 ne produit pas de probabilités associées (voir 2 en marge gauche). Le dialogue, réalisé sous CMS, est reproduit ci-dessous.

```
     Quel est le nom du FICHIER DE DONNEES? (Par defaut: "... DATA A")
     lacs donnees a
     Quel doit etre le nom du FICHIER DE SORTIE contenant la matrice
     de ressemblance? (Par defaut: "... DATA A")
     lacs s15 a
     Quel est le nom du fichier contenant les MATRICES DE SIMILARITES
     PARTIELLES, si ce probleme en comporte? (Par defaut: "... DATA A")
(1)
     Quel doit etre le nom du fichier contenant la MATRICE DES
     PROBABILITES, si vous desirez l'obtenir?
     (Par defaut: "PROBAB DATA A")
(2)
```

```
 S I M I L : Calcul de matrices de ressemblance.

 VERSION 3.0b
 AUTEUR: A. VAUDOR
 REFERENCE: Chapitre 7 de
 Legendre, L. et P. Legendre.  1984 -- Ecologie numerique,
     2ieme edition. Tome 2: La structure des donnees ecol-
     giques. Collection d'Ecologie, 13.  Masson, Paris  et
     les Presses de l'Universite du Quebec.  viii + 335 p.

 TITRE:
physico-chimie  de  71  lacs.
 NOMBRE D'OBJETS (LIGNES)?
71
 NOMBRE DE DESCRIPTEURS (COLONNES)?
11
 CODE REPRESENTANT L'ABSENCE D'INFORMATION?
-9
 LES 10 PREMIERES COLONNES CONTIENNENT-ELLES LES NOMS DES OBJETS? (o ou n)
o
 QUELLE FONCTION DESIREZ-VOUS CALCULER ?
 Similarites   : s01  a  s26, ou  Nei
 Distances     : d01  a  d14
 Mode R:    rp = r de Pearson
            rs = r de Spearman
            tau= Tau de Kendall
            khi= Khi-carre (statistique G)
            ht = Coefficient de contingence de Tschuproff
            hs0= Information reciproque         S=B/(A+B+C)
            hs1= Coherence de Rajski     S'=SQRT(1-(hd)**2)
            hs2= Coeff. symetr. d'incertitude S"=B/(A+2B+C)
            hd = Metrique de Rajski        D =(A+C)/(A+B+C)
s15
 LES ECARTS MAXIMA DES VARIABLES (R[i])
 DOIVENT-ILS ETRE CALCULES A PARTIR DES DONNEES? (o ou n)
o
 LES POIDS W[i] DOIVENT-ILS ETRE SIMPLEMENT 0 OU 1? (o ou n)
o
 Combien y a-t-il de descripteurs QUALITATIFS multi-classes?
0
 Fin du programme.
```

Contenu du fichier de résultats

Le fichier de sortie contenant les résultats des calculs est écrit en binaire; il est donc impossible de le consulter directement à l'aide d'un éditeur de texte. Il en est de même du fichier des probabilités associées à certains coefficients. On peut cependant consulter ces fichiers grâce au programme utilitaire REGARDE qui les transcrira en caractères lisibles.

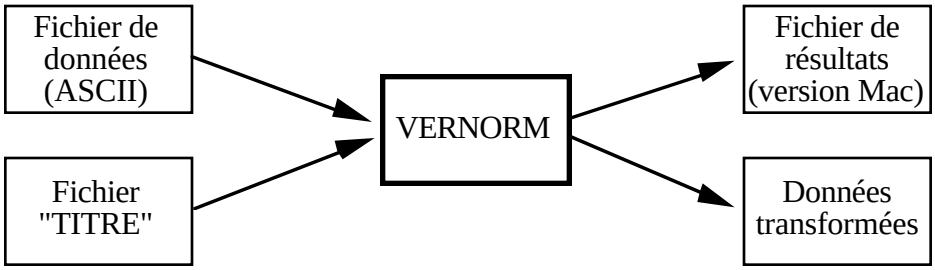
VERNORM

Que fait VERNORM ?

VERNORM est un programme à multiples usages destiné à être utilisé en début de traitement d'un tableau de données. Ce programme a été créé pour répondre à des besoins généraux concernant les fichiers d'entrée. Son nom signifie VÉRifier et NORMaliser; il peut aussi transposer une matrice, reformater les données, enlever ou ajouter des identificateurs d'objets, dessiner des histogrammes, diviser les variables en classes, centrer et réduire les données ou les transformer de diverses façons. VERNORM sait tenir compte des absences d'information. L'élimination de lignes et de colonnes d'un fichier de données pourra être réalisée à l'aide des programmes de bases de données ou des éditeurs disponibles sur micro-ordinateurs; de même, des programmes statistiques pourront être employés pour réaliser certaines transformations de données non disponibles dans VERNORM.

Afin d'effectuer l'opération désirée, il suffit simplement de choisir l'option voulu et de répondre aux questions posées par VERNORM. Laissez-vous guider par le programme.

Fichiers d'entrée et de sortie



(1) Fichier de données brutes

Ce fichier, écrit en ASCII (caractères lisibles), se présente sous la forme d'un tableau $p \times n$ de données brutes où les lignes sont habituellement les p objets et les colonnes sont les n variables. Il peut contenir des identificateurs dans les 10 premières colonnes, si l'usager le désire; son format général est présenté à la section "Fichier d'entrée principal" dans la description du programme SIMIL. Il peut avoir été tapé à l'aide d'un éditeur, ou encore tiré d'un chiffrier (en format "texte"). Il peut également être le résultat de calculs réalisés par un logiciel statistique (SPSS, SAS, STATVIEW, etc.) ou par un autre programme capable d'écrire ses résultats sur un fichier "texte" (EBCDIC ou ASCII).

(2) Fichier d'identificateurs

Si le tableau de données ne comporte pas d'identificateurs d'objets, ou encore si on l'a transposé, il est possible d'ajouter des noms aux lignes en fournissant à VERNORM un fichier contenant une liste d'identificateurs de lignes. Dans les versions CMS et VMS, ce fichier porte par défaut le nom de "TITRE" et se présente sous la forme d'une liste de noms d'objets inscrits dans un fichier lisible (EBCDIC ou ASCII).

Attention : Sur Macintosh, ces identificateurs doivent posséder au minimum 10 caractères, incluant les espaces et les tabulateurs; les 10 premiers caractères de chaque ligne seront utilisés. S'il y en a moins et si on n'a pas complété à l'aide de blancs, le programme complétera en utilisant des caractères de la ligne suivante, si bien qu'il manquera d'identificateurs en fin de liste. Dans les versions CMS et VMS, il suffit d'inscrire les identificateurs sur les lignes successives du fichier; les 10 premiers caractères de chaque ligne seront utilisés. Exemple d'un fichier de noms d'espèces de kangourous:

```
Setonix b.
Thylog. c.
Petrog. g.
Wallab. v.
Macrop. w.
```

(3) Fichier de données transformées

Après transformation du fichier, les données peuvent être inscrites dans un nouveau fichier si l'usager en fait la demande. Ce fichier, d'une largeur n'excédant pas 80 caractères, est d'un format approprié pour servir d'entrée au programme SIMIL.

(4) Fichier de résultats statistiques

Dans la version Macintosh, les résultats des calculs de VERNORM sont présentés dans un fichier contenant des informations statistiques diverses, selon les opérations qui ont été demandées. Son contenu est décrit plus en détail à la section "Contenu du fichier de résultats". Dans les versions pour grands ordinateurs, ces mêmes informations sont présentées à l'écran (voir l'exemple); elles pourront être conservées dans un fichier de mémoire de console, en suivant les instructions de la p. 2.

Les options du programme

Ce programme offre une large gamme d'options. Les options préliminaires permettent de vérifier les données, de transposer la matrice de départ ou de rendre les données positives; les options principales, numérotées de 0 à 8 dans les versions CMS et VMS, permettent de tester la normalité des données, de les transformer de plusieurs façons et de tracer des histogrammes. Les détails d'utilisation des différentes options décrites ci-dessous se trouvent à la section suivante ("Les questions du programme"). Voyons ces options une à une.

(1) Vérification du fichier d'entrée

Cette option permet d'abord de s'assurer que le fichier est complet et qu'il comporte le bon nombre de lignes et de colonnes. S'il y a moins de données que déclaré par l'usager, ou encore si des nombres sont accolés (pas d'espace) ou contiennent des caractères "illégaux", un message d'erreur est émis et le programme s'arrête. Cette vérification fournit également les bornes globales des valeurs dans le tableau (minimum, maximum), ainsi que le nombre de valeurs et les bornes par ligne, toutes informations utiles pour détecter les problèmes; voir l'exemple.

(2) Transposition de la matrice de données

Cette option offre la possibilité de transposer la matrice $p \times n$ de départ en une matrice $n \times p$, ou l'inverse. Ainsi on pourra, à partir du même fichier, avoir accès à des analyses en mode Q et en mode R en transposant la matrice initiale. En ayant recours à un "Fichier d'identificateurs" (voir plus haut), on peut munir les lignes du fichier transposé d'identificateurs (10 caractères) au moment de la réécriture des données.

(3) Rendre les données positives

Cette option permet d'éliminer les valeurs nulles ou négatives du fichier, par addition d'une constante différente pour chaque colonne, ou encore d'une valeur unique pour tout le fichier. Cette translation est obligatoire si on désire utiliser la transformation de Taylor, de Box-Cox ou de Box-Cox-Bartlett, car ces transformations requièrent le calcul de logarithmes des données; elle peut également être utile comme préliminaire au calcul de certains indices de similarité.

(4) OPTION 0: Transformation de Taylor

Cette transformation a pour but premier d'homogénéiser les variances des variables de la matrice initiale. Il faut préalablement avoir rendu les données strictement positives pour utiliser cette transformation, car elle requiert le calcul des logarithmes des données. Lorsque le jeu de données comprend plusieurs groupes d'objets (formant les colonnes du tableau), ou encore lorsqu'on traite un groupe de descripteurs quantitatifs dimensionnellement homogènes (ex.: abondances d'espèces) auxquels on désire faire subir la même transformation, la loi des puissances de Taylor fournit une transformation générale qui tend à rendre les variances homogènes. Les données ont ainsi plus de chances de se conformer aux conditions requises par la statistique paramétrique, ce qui inclut la normalité. Si on trace un graphique de la variance de chacune des variables considérées par rapport à sa moyenne, la loi de Taylor relie les moyennes aux variances par l'équation

$$\mathrm{Var}(y) = a\,(\mathrm{Moy}(y))^b$$

qui permet de calculer la valeur des paramètres a et b par régression non-linéaire. On peut aussi en calculer une approximation par régression linéaire (modèle I ou modèle II) sur la forme logarithmique de la même équation. VERNORM offre les options suivantes pour le calcul de cette régression:

- Modèle I: régression linéaire simple.
- Modèle II: méthode de l'axe majeur réduit.
- Modèle II: méthode des trois groupes de Bartlett.
- Modèle II: méthode de l'axe principal.
- Régression non-linéaire.

On peut également demander au programme de calculer toutes les solutions ci-dessus. Les différences entre ces méthodes sont exposées à la section 4.3 de Legendre & Legendre (1984a) ainsi que dans de nombreux manuels de statistique.

(5) OPTION 1: Transformation de Box et Cox

Cette option permet de normaliser individuellement les variables de la matrice de données. La méthode de Box-Cox permet de déterminer empiriquement quel est l'exposant produisant une distribution qui s'approche le plus d'une normale, dans la fonction générale de la transformation

$$y' = (y^{lambda} - 1)/lambda \qquad \text{si } lambda \neq 0$$

et

$$y' = \ln(y) \qquad \text{si } lambda = 0.$$

La valeur de *lambda* est trouvée en maximisant une fonction de vraisemblance, par recherche itérative (Sokal & Rohlf, 1981: 423). Toutes les valeurs de y doivent être strictement positives car la fonction de vraisemblance requiert le calcul des logarithmes des données.

Lorsque *lambda* est égal à 1, la fonction serait une simple transformation linéaire; en pratique, aucune transformation n'est effectuée dans ce cas. Si *lambda* est égal à 0.5, la fonction produit la transformation √; lorsque *lambda* est égal à 0, la transformation est log; enfin, lorsque *lambda* est égal à -1, on obtient la transformation inverse. Cette méthode, très efficace pour réduire l'asymétrie des données, ne saurait en aucun cas rendre normale une distribution possédant plusieurs modes.

(6) OPTION 2: Transformation de Box-Cox-Bartlett

Cette option permet également de normaliser les variables tout en homogénéisant leurs variances. Dans cette variante, on utilise la statistique χ^2 d'homogénéité des variances de Bartlett dans l'équation du maximum de vraisemblance de la méthode de Box-Cox (Sokal & Rohlf, 1981: 425); elle conduit à une transformation unique, pour toutes les variables du fichier, qui homogénéise au mieux les

variances tout en normalisant les distributions. Comme la transformations de Taylor, cette option est utilisable dans le cas où on désire faire subir la même transformation à tout un groupe de variables quantitatives. Toutes les valeurs de *y* doivent être strictement positives car la fonction de vraisemblance requiert le calcul des logarithmes des données.

(7) OPTION 3: Division en classes

Cette option offre à l'utilisateur la possibilité de diviser les variables du fichier d'entrée en classes. Il peut choisir de diviser toutes les variables ou seulement certaines de celles-ci. Il peut de plus décider d'un nombre de classes pour chaque variable, ou encore il peut décider de diviser toutes les variables en un même nombre de classes. VERNORM proposera un nombre de classes *k* en fonction du nombre d'observations *p*, selon la règle de Sturge: $k = 1 + (3.3 \log_{10} p)$ avec arrondi de *k* à la valeur entière la plus proche.

(8) OPTION 4: Votre choix de transformations

Cette option permet de choisir entre quatre familles de transformations:

1) $y' = a + by$
2) $y' = y^a$
3) $y' = \exp(y)$
4) $y' = \ln(a + by)$

Le cas échéant, l'usager devra indiquer les valeurs des constantes *a* et *b*. Cette option permet de transformer toutes les variables du fichier, ou certaines seulement. Notons qu'il est souvent plus facile de réaliser ces transformations à l'intérieur de chiffriers (ex. EXCEL) ou de logiciels statistiques (ex. STATVIEW) disponibles sur microordinateurs.

(9) OPTION 5: Histogrammes

Cette option dessine les histogrammes de fréquence pour toutes les variables. Ceci permet de visualiser graphiquement les distributions des descripteurs avant de choisir comment les transformer. Dans les versions CMS et VMS, l'histogramme est représenté latéralement à l'écran (voir l'exemple), alors que dans la version Macintosh, il s'agit d'un histogramme dessiné de la façon habituelle (voir le "Contenu du fichier de résultats"). Le nombre de classes est déterminé par l'usager; VERNORM lui propose un nombre de classes *k* en fonction du nombre d'observations *p*, selon la règle de Sturge: $k = 1 + (3.3 \log_{10} p)$ avec arrondi de *k* à la valeur entière la plus proche.

(10) OPTION 6: Centrage et réduction

Cette option permet de centrer et réduire les variables choisies par l'utilisateur ("*z*-scores" en anglais). En utilisant cette transformation après une normalisation des données, on obtient des variables normales centrées-réduites.

(11) OPTION 7: Tests de normalité

Cette option calcule le test de normalité Kolmogorov-Smirnov en se référant à la table des valeurs critiques proposée par Lilliefors (1967); cette table tient compte du fait que la moyenne et la variance de la population ne sont pas connues par hypothèse mais sont plutôt estimées à partir des données elles-mêmes. Le test de Kolmogorov-Smirnov est préférable au test khi-carré par exemple, car ce dernier ne prend pas en compte la nature ordonnée des données. Le calcul est réalisé pour toutes les variables du fichier d'entrée. Les résultats sont fournis pour le seuil de signification choisi par l'utilisateur; on se rappellera qu'un seuil de signification plus bas (1% par exemple) est plus permissif quant à la distribution des données, car il est alors plus difficile de rejeter l'hypothèse de normalité.

(12) OPTION 8: Réécriture du fichier des données transformées

Cette option est utilisée après transformation des variables ou transposition de la matrice, pour récrire les données transformées dans un nouveau fichier; les données récrites auront subi toutes les transformations réalisées jusqu'alors. Une série de questions sont posées par le programme pour déterminer la structure d'écriture des données dans le fichier. Il est possible à cette étape de recoder l'absence d'information, d'imposer une échelle aux données en fixant le minimum et le maximum et d'inclure des identificateurs de lignes, en fournissant un fichier d'identificateurs.

Les questions du programme

Les questions présentées par le programme à l'écran du Macintosh sont décrites dans les paragraphes qui suivent. Les questions posées par les versions CMS et VMS sont essentiellement les mêmes, comme on peut le vérifier avec l'exemple présenté à la section suivante. Pour faire démarrer le programme, il faut cliquer sur l'icône, puis donner la commande "Ouvrir" dans le menu "Fichiers".

(1) "Fichier d'entrée" — Le programme présente la liste des fichiers ASCII disponibles.

(2) "Fichier de résultats statistiques" — Le programme présente le menu permettant de donner un nom au fichier de résultats statistiques calculés par le programme. Cette question n'est pas posée par les versions CMS et VMS car les résultats statistiques apparaissent à l'écran.

(3) "Le fichier est-il identifié (10 premiers caractères)? [Oui, Non] — On répond *Oui* si les 10 premières colonnes de chaque vecteur-objet ou vecteur-descripteur contiennent des identificateurs.

(4) "Nombre de lignes (ou de blocs de lignes)" — La réponse doit être un nombre entier positif. Dans le cas d'une matrice p lignes (objets) x n colonnes (variables), on demande ici le nombre d'objets, chaque objet pouvant occuper une ou plusieurs lignes; si la matrice est transposée, il s'agit du nombre de variables, chaque variable pouvant également occuper une ou plusieurs lignes; voir la description du fichier de données brutes.

(5) "Nombre de colonnes" — Dans le cas d'une matrice p lignes (objets) x n colonnes (variables), on inscrit le nombre de descripteurs décrivant chaque objet du fichier, à l'exclusion des identificateurs de lignes si le fichier en possède. Si le fichier est transposé, on inscrit le nombre d'objets composant chaque vecteur-variable, à l'exclusion des identificateurs de descripteurs.

(6) "Valeur indiquant l'absence d'information" — On inscrit quelle valeur numérique a été utilisée dans le fichier pour indiquer qu'une information est absente (souvent: -1, -9, -999, etc.). On doit répondre à cette question par une valeur numérique même s'il n'y a pas de données manquantes dans le fichier.

(7) "Vérification du fichier d'entrée?" [Oui, Non] — Voir la description de cette fonction au paragraphe (1) des options. Si on répond *Oui*, le programme demande des informations additionnelles quant au contenu du fichier d'entrée:

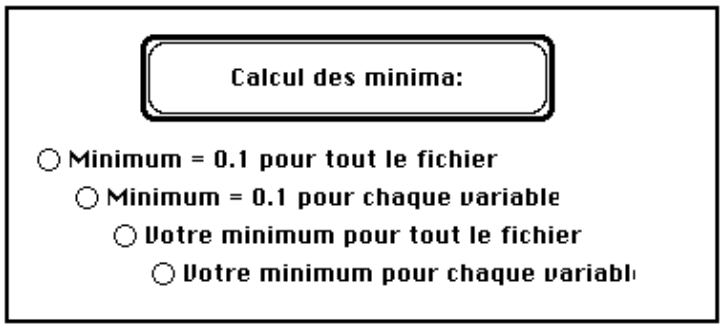
(7.1) "Fichier d'entrée ne contenant que des entiers?" [Oui, Non] — Selon la nature numérique des données, entières ou réelles, le programme emploie des procédures différentes pour lire les données. Après la réponse à cette question, le programme (versions Macintosh aussi bien que grands ordinateurs) liste à l'écran le nombre de valeurs ainsi que le minimum et le maximum de chaque ligne; voir l'exemple. À la fin de cette liste, le programme fournit la plus petite et la plus grande valeur du fichier; cliquez la souris pour obtenir la question suivante.

(8) "Transposition de la matrice de données?" [Oui, Non] — On répond *Oui* si on désire que la matrice de données soit transposée. Les identificateurs de lignes, qui deviennent les colonnes au cours de cette

opération, sont perdus. Une nouvelle série de noms, fournis dans un fichier d'identificateurs, pourront être ajoutés au début des nouvelles lignes si l'usager le désire.

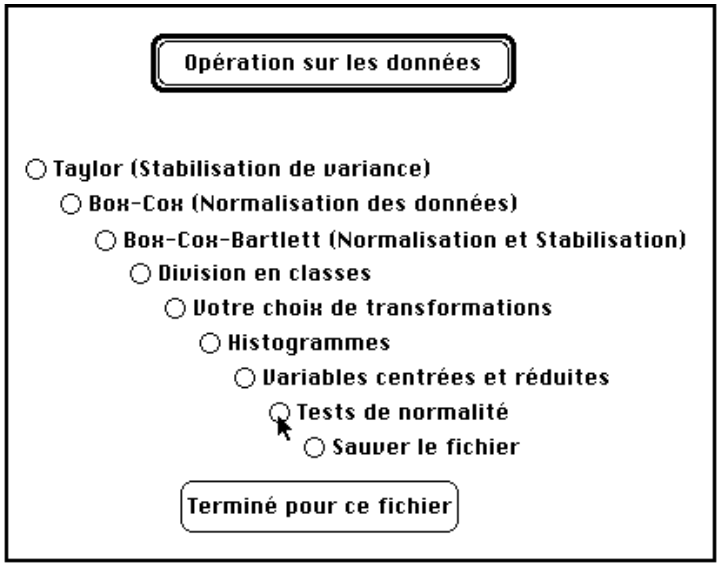
(9) "Rendre les données positives (plus particulièrement pour Taylor, Box-Cox, Box-Cox-Bartlett) ?" [Oui, Non] — Voir la description de cette fonction au paragraphe (3) des options. Si on répond *Oui*, le programme présente l'écran d'options suivant:

(9.1)



Comme on le voit, l'usager peut décider d'imposer la valeur minimum qu'il désire (une valeur pour chaque variable, ou encore une valeur unique pour tout le fichier), ou il peut charger le programme d'imposer la valeur minimum de 0.1 (séparément pour chaque variable, ou encore comme si le fichier ne contenait qu'une seule variable).

(10) "Opération sur les données" — L'écran permettant à l'usager de choisir l'opération désirée se présente comme suit:

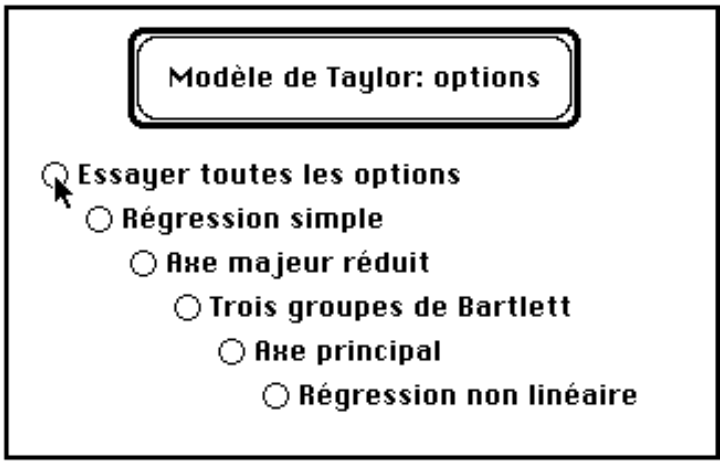


Dans les versions CMS et VMS, cet écran est divisé en deux questions. D'abord, "Désirez-vous effectuer une manipulation sur le fichier? (O ou N)". Si on répond *Non*, le programme s'arrête immédiatement; si on répond *Oui*, le menu des 9 options ci-dessous est présenté; voir l'exemple.

À partir de ce point, les questions divergent selon l'option choisie. Après le calcul de chaque option, le menu ci-dessus est présenté de nouveau. Pour arrêter le cycle, il faut pousser le bouton "Terminé pour ce fichier"; on retourne alors au menu "Fichiers" qui permet de traiter un autre fichier de données. La commande "Interrompre" dans le menu "R: Vernorm" permet de quitter le programme.

Bouton: Taylor (Stabilisation de la variance) — Voir le paragraphe (4) des options

(11) “Modèle de Taylor: options” — Le programme présente l’écran suivant:



Bouton: Box-Cox (Normalisation des données) — Voir le paragraphe (5) des options

(11) “Combien de variables à transformer?” — On indique le nombre de variables devant subir la transformation de Box & Cox. Si on demande de transformer toutes les variables, aucune autre question n’est posée. Dans le cas contraire, le programme demande:

(11.1) “Numéro de la variable [1]” — Si par exemple la cinquième variable du fichier est la première devant être transformée, on répond “5”. Le programme demande ensuite: “Numéro de la variable [2]”, etc. jusqu’à épuisement du nombre de variables à transformer.

Bouton: Box-Cox-Bartlett (Normalisation et Stabilisation) — Paragraphe (6) des options

Aucune question additionnelle n’est posée par le programme. Une transformation unique est calculée pour l’ensemble des variables. Cette transformation n’est pas la même que celle qui aurait été obtenue si on avait demandé la transformation de Box-Cox ci-dessus (paragraphe 5 des options) après avoir déclaré tout le fichier comme formant une seule variable.

Bouton: Division en classes — Voir le paragraphe (7) des options

(11) “Combien de variables à transformer?” — On indique le nombre de variables devant subir la division en classes. Si on n’ordonne pas de transformer toutes les variables, le programme demande:

(11.1) “Numéro de la variable [1]” — Si par exemple la cinquième variable du fichier est la première devant être divisée en classes, on répond “5”. Le programme demande ensuite: “Numéro de la variable [2]”, etc. jusqu’à épuisement du nombre de variables à diviser.

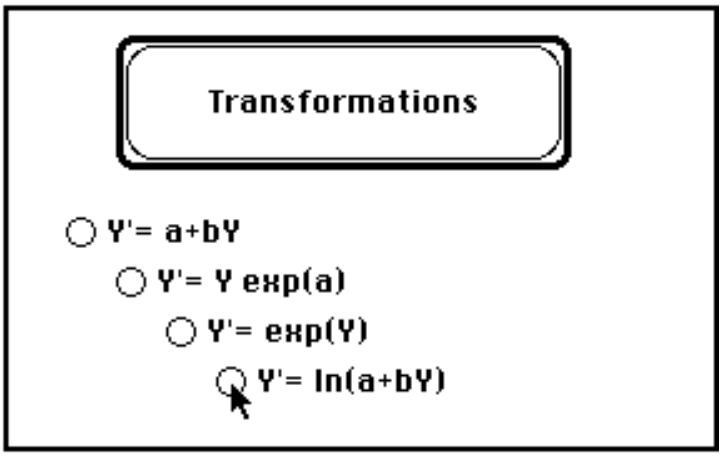
(12) “Même nombre de classes pour chacune des variables?” [Oui, Non] — On répond *Non* si on désire diviser chaque variable en un nombre différent de classes.

(12.1) Si on a répondu *Oui* en (12), le programme pose la question suivante: “Nombre de classes? (Nombre de Sturge = k)” — On indique le nombre de classes désiré.

(12.2) Si on a répondu *Non* en (12), le programme pose la question suivante: “Nombre de classes de la variable [1]? (Nombre de Sturge = k)” — On indique le nombre de classes désiré pour la première variable. Le programme répète la question pour les autres variables identifiées à la question (11.1).

Bouton: Votre choix de transformations — Voir le paragraphe (8) des options

(11) “Transformations” — Le programme présente l’écran suivant:



(12) “Valeur de a” et/ou “Valeur de b” — Selon la réponse en (11), le programme demande la valeur des paramètres de la transformation désirée. Pour la transformation classique des abondances d’espèces, $y' = \ln(y + 1)$, par exemple, il suffit de presser le dernier bouton et de donner les valeurs suivantes pour les paramètres: $a = 1$, $b = 1$.

(13) “Combien de variables à transformer?” — On indique le nombre de variables devant subir la transformation choisie. Si on n’ordonne pas de transformer toutes les variables, le programme demande:

(13.1) “Numéro de la variable [1]” — Si par exemple la cinquième variable du fichier est la première devant être transformée, on répond “5”. Le programme demande ensuite: “Numéro de la variable [2]”, etc. jusqu’à épuisement du nombre de variables à transformer.

Bouton: Histogrammes — Voir le paragraphe (9) des options

(11) “Nombre de classes? (Nombre de Sturge = *k*) — On indique le nombre de classes désiré. Les histogrammes apparaissent à l’écran; il faut “terminer” chaque graphique pour obtenir le suivant. Les dessins peuvent également être imprimés ou conservés sur un fichier de type PICT pour utilisation future.

Bouton: Variables centrées et réduites — Voir le paragraphe (10) des options

(11) “Combien de variables à transformer?” — On indique le nombre de variables devant subir cette transformation. Si on n’ordonne pas de transformer toutes les variables, le programme demande:

(11.1) “Numéro de la variable [1]” — Si par exemple la cinquième variable du fichier est la première devant être transformée, on répond “5”. Le programme demande ensuite: “Numéro de la variable [2]”, etc. jusqu’à épuisement du nombre de variables à transformer.

Bouton: Tests de normalité — Voir le paragraphe (11) des options

(11) “Tests de normalité de Kolmogorov-Smirnov-Lilliefors. Seuil de signification:” [1%, 5%, 10%, 15%, 20%] — On indique le niveau de signification désiré en appuyant sur le bouton correspondant.

Bouton: Sauver le fichier — Voir le paragraphe (12) des options

(11) “Nombre de caractères dans lesquels les nombres seront écrits” — On indique, en nombre de caractères, la largeur du champ qui doit être consacré à chaque variable. Voir l’exemple.

(12) “Nombre de chiffres après le point” — On indique le nombre de décimales qui seront retenues pour chaque variable. Voir l’exemple.

(13) “Remplacement de l’absence d’information par” — On indique le code de l’absence d’information dans le fichier transformé. Ce code peut être le même que dans le fichier d’entrée, ou non. On doit fournir une valeur numérique même s’il n’y a pas de données manquantes dans le fichier.

(14) “Désirez-vous fixer les min et max du fichier de sortie?” [Oui, Non] — On répond *Non* si on ne désire pas imposer une échelle aux données en fixant leur minimum et leur maximum. Si on répond plutôt *Oui*, le programme pose les questions suivantes:

(14.1) “Valeur minimum” — La valeur fournie sert de borne minimum pour l’ensemble du fichier.

(14.2) “Valeur maximum” — La valeur fournie sert de borne maximum pour l’ensemble du fichier.

(15) “Existe-t-il un fichier d’identificateurs?” [Oui, Non] — On répond *Oui* si on a préparé un fichier d’identificateurs de lignes; le programme présente alors le menu des fichiers ASCII disponibles. Les identificateurs seront transcrits dans les 10 premières colonnes de chaque ligne (ou bloc de lignes).

Exemple

Voici un exemple d’utilisation du programme VERNORM sur grands ordinateurs. Le fichier de données contient 60 objets et 3 descripteurs. Même s’il n’y a pas d’absence d’information dans ce fichier, on doit quand même fournir une réponse à cette question (les réponses de l’usager sont en souligné gras). On demande d’abord au programme de calculer des tests de normalité de Kolmogorov-Smirnov sur les données brutes et de tracer des histogrammes; on demande ensuite de rechercher la meilleure transformation normalisatrice suivant Box & Cox, puis de réaliser à nouveau des tests de normalité et de tracer des histogrammes pour les variables transformées. Enfin, on demande de récrire les données dans un format de 10 caractères avec 5 décimales (format Fortran 3F10.5). Le dialogue, réalisé sous CMS, est reproduit ci-dessous. Le fichier traité est le même que dans l’exemple du “Contenu du fichier de résultats” ci-dessous, où il est analysé sur Macintosh.

```
Vernorm
Quel est le nom du FICHIER DE DONNEES? (Par defaut: "... data a")
60x3  data  a

Quel doit etre le nom du fichier de DONNEES TRANSFORMEES?
(Par defaut: "... data a")
60x3  transfor  a

Quel est le nom du fichier contenant les NOMS DES OBJETS, s'il
y a lieu? (Par defaut: "TITRES data a")

Execution begins...
 P R O G R A M M E   V E R N O R M
 pour VERifier et NORMaliser les tableaux de donnees
 VERSION 3.0b
 AUTEUR: A. VAUDOR.
```

```
 LES 10 PREMIERES COLONNES CONTIENNENT-ELLES LES NOMS DES OBJETS? (o ou n)
n
 NOMBRE D'OBJETS (LIGNES)?
60
 NOMBRE DE DESCRIPTEURS (COLONNES)?
3
 CODE DESIGNANT L'ABSENCE D'INFORMATION?
-999
 DESIREZ-VOUS EFFECTUER LA VERIFICATION DU FICHIER DE DONNEES?
o
 S'AGIT-IL D'UN FICHIER DE DONNEES NE CONTENANT QUE DES ENTIERS?
n
 LIGNE    N. DE VALEURS      MIN         MAX
     1             3        3.08       48.70
     2             3        2.84       48.20
     3             3        3.12       49.00
     4             3        3.37       48.40
  [etc.]
    59             3        2.90       42.20
    60             3        0.86       42.20

 PLUS PETITE VALEUR DANS LE FICHIER:    0.23
 PLUS GRANDE VALEUR DANS LE FICHIER:   50.10

 DESIREZ-VOUS TRANSPOSER LE FICHIER DE DONNEES?
n
 DESIREZ-VOUS RENDRE TOUTES LES DONNEES POSITIVES?
(CECI EST NECESSAIRE POUR LES TRANF. DE TAYLOR, BOX-COX ET BOX-COX-BARTLETT)
n
 DESIREZ-VOUS EFFECTUER UNE MANIPULATION SUR LE FICHIER? (O ou N)?
o
  OPTIONS
    0: TAYLOR (homogeneise les variances)
    1: BOX-COX (normalise les donnees)
    2: BOX-COX-BARTLETT (normalise les donnees ET homogeneise les variances)
    3: DIVISION EN CLASSES
    4: VOTRE CHOIX DE TRANSFORMATION
    5: HISTOGRAMMES
    6: VARIABLES CENTREES REDUITES)
    7: TESTS DE NORMALITE Kolmogorov-Smirnov-Lilliefors
    8: REECRITURE DU FICHIER DE DONNEES
7
 SEUIL DE SIGNIFICATION DESIRE: INSCRIVEZ
     1 = 1 %, 2 = 5 %, 3 = 10 %, 4 = 15 %, 5 = 20 %
2

 TEST DE KOLMOGOROV-SMIRNOV (TABLE DE LILLIEFORS )

 HYPOTHESE : R=REJETEE, NR=NON REJETEE, NC=NON CALCULABLE

 VARIABLE  :    1      2      3
 DISTANCE  : 0.1667 0.2629 0.0821
 VAL. CRIT.: 0.1144 0.1144 0.1144
 HYPOTHESE :     R      R     NR
```

```
 DESIREZ-VOUS EFFECTUER UNE MANIPULATION SUR LE FICHIER? (O ou N)?
o
  OPTIONS
    0: TAYLOR (homogeneise les variances)
    1: BOX-COX (normalise les donnees)
    2: BOX-COX-BARTLETT (normalise les donnees ET homogeneise les variances)
    3: DIVISION EN CLASSES
    4: VOTRE CHOIX DE TRANSFORMATION
    5: HISTOGRAMMES
    6: VARIABLES CENTREES REDUITES)
    7: TESTS DE NORMALITE Kolmogorov-Smirnov-Lilliefors
    8: REECRITURE DU FICHIER DE DONNEES
5

 HISTOGRAMMES

 NOMBRE DE CLASSES? (LA REGLE DE STURGE SUGGERE 7 CLASSES; MAX = 60 CLASSES)
7

 VARIABLE :  1
 |*************
 |*************************
 |************
 |****
 |**
 |***
 |*
 ------------------------------------------------------------------------------
                         25

 VARIABLE :  2
 |**
 |**************
 |****
 |
 |*
 |********************
 |********************
 ------------------------------------------------------------------------------
                     20

 VARIABLE :  3
 |***
 |*****
 |*************
 |********************
 |****************
 |**
 |**
 ------------------------------------------------------------------------------
                    19

 DESIREZ-VOUS EFFECTUER UNE MANIPULATION SUR LE FICHIER? (O ou N)?
o
```

```
  OPTIONS
    0: TAYLOR (homogeneise les variances)
    1: BOX-COX (normalise les donnees)
    2: BOX-COX-BARTLETT (normalise les donnees ET homogeneise les variances)
    3: DIVISION EN CLASSES
    4: VOTRE CHOIX DE TRANSFORMATION
    5: HISTOGRAMMES
    6: VARIABLES CENTREES REDUITES)
    7: TESTS DE NORMALITE Kolmogorov-Smirnov-Lilliefors
    8: REECRITURE DU FICHIER DE DONNEES
1
 COMBIEN DE VARIABLES VOULEZ-VOUS TRANSFORMER ?
3

 TRANSFORMATION DE BOX ET COX

    VARIABLE    1
 LIMITE DE L'I.C. DE LAMBDA  -6.31388
                     LAMBDA  -3.96688
 LIMITE DE L'I.C. DE LAMBDA  -1.78488
    VARIABLE    2
 LIMITE DE L'I.C. DE LAMBDA   3.62280
                     LAMBDA   8.77780
 LIMITE DE L'I.C. DE LAMBDA  14.08280
    VARIABLE    3
 LIMITE DE L'I.C. DE LAMBDA   0.59872
                     LAMBDA   1.03672
 LIMITE DE L'I.C. DE LAMBDA   1.52372

 DESIREZ-VOUS EFFECTUER UNE MANIPULATION SUR LE FICHIER? (O ou N)?
o
  OPTIONS
    0: TAYLOR (homogeneise les variances)
    1: BOX-COX (normalise les donnees)
    2: BOX-COX-BARTLETT (normalise les donnees ET homogeneise les variances)
    3: DIVISION EN CLASSES
    4: VOTRE CHOIX DE TRANSFORMATION
    5: HISTOGRAMMES
    6: VARIABLES CENTREES REDUITES)
    7: TESTS DE NORMALITE Kolmogorov-Smirnov-Lilliefors
    8: REECRITURE DU FICHIER DE DONNEES
8
 NOMBRE D'ESPACES DANS LESQUELS LES NOMBRES SERONT ECRITS?
10
 COMBIEN DE DECIMALES APRES LE POINT?
5
 CODE DESIRE POUR LES VALEURS ABSENTES:
-999
 DESIREZ-VOUS IMPOSER UNE ECHELLE AUX VALEURS
 EN FIXANT VOUS-MEME LE MINIMUM ET LE MAXIMUM?
n
 AVEZ-VOUS PREPARE UN FICHIER DE NOMS D'OBJETS? (Fichier "TITRE")?
n
```

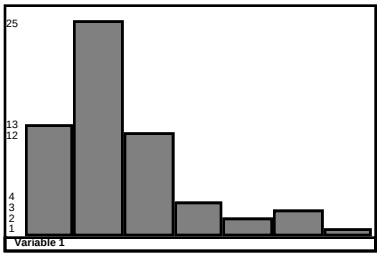
```
TRANSFORMATIONS EFFECTUEES SUR LE FICHIER DE SORTIE:
0=pas de transformation, A=Box-Cox, B=division en classes
C=votre choix de transformation, D=var. centree reduite
AA0                                              [Explication: section suivante]

CHAMP INSUFFISANT; NOUVEAU FORMAT: 3F22.5         [Explication: section suivante]
Fin du programme.
```

Contenu du fichier de résultats

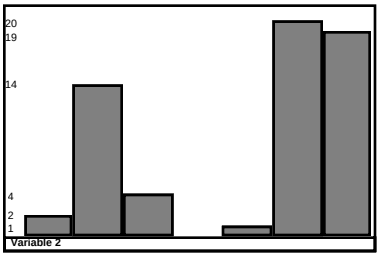
Le fichier de résultats statistiques produit par la version Macintosh contient différentes informations, selon les options du programme qui ont été choisies. Voici un exemple de contenu d'un tel fichier; les commentaires sont intercalés parmi les résultats. Ces mêmes informations apparaissent à l'écran dans les versions pour grands ordinateurs; voir ci-dessus.

(1) Comme dans l'exemple de la section précédente (mêmes données), on a d'abord demandé de calculer des tests de normalité de Kolmogorov-Smirnov, au seuil de signification de 5%. Les histogrammes correspondants, qui étaient présentés à l'écran, ont été repiqués sur un fichier PICT et recopiés ci-dessous; la fréquence des différentes colonnes est indiquée à gauche, alors que le numéro de la variable est inscrit au bas de chaque graphique (texte trop petit pour être lu sur ces copies fortement réduites). Le test rejette l'hypothèse de normalité pour les variables 1 et 2. Les histogrammes en montrent les raisons: la distribution de la variable 1 est fortement asymétrique vers la droite et devrait être normalisable; la variable 2 présente une distribution bimodale qui n'est pas normalisable par une transformation telle que celles proposées dans ce programme; enfin, la variable 3 est déjà reconnue comme normale par le test K-S et présente une distribution unimodale et symétrique.

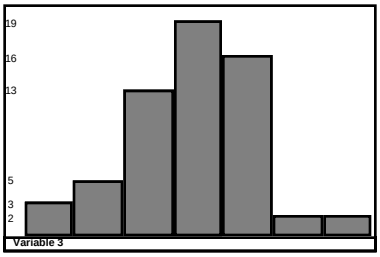
```
Tests de Kolmogorov-Smirnov-Lilliefors

 Hypothèses: r=Rejetée, Nr=non rejetée, Nc=non calculable

   Variable:    1      2      3
   Distance: 0.1667 0.2629 0.0821
   Val.Crit: 0.1144 0.1144 0.1144
 Hypothèse :      R      R     Nr
```



Variable 1



Variable 2



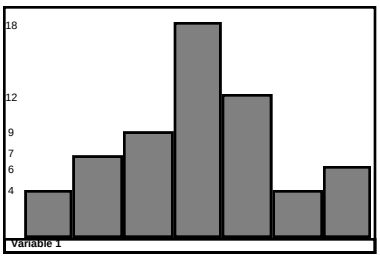
Variable 3

(2) On demande ensuite au programme de calculer la meilleure transformation normalisatrice suivant Box & Cox. Pour chaque variable, le programme présente la valeur de maximum de vraisemblance du paramètre *lambda*, accompagnée des bornes de l'intervalle de confiance de 95% ("limite lambda"). Pour la première variable, la valeur -3.96688 sera employée comme exposant dans la transformation de Box-Cox; pour la variable 2, c'est la valeur 8.77780 qui sera employée, puisque la valeur "1" ne se trouve pas à l'intérieur de l'intervalle de confiance du paramètre. Quant à la troisième variable, quoique la meilleure valeur du paramètre soit 1.03672, aucune transformation ne sera faite puisque la valeur "1" (pas de transformation) se trouve à l'intérieur de l'intervalle de confiance de 95%.

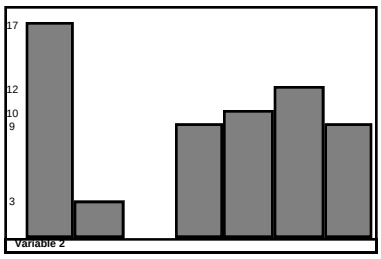
```
Transformations de Box et Cox

      Variable    1
        limite lambda  -6.31388
               lambda  -3.96688
        limite lambda  -1.78488
      Variable    2
        limite lambda   3.62280
               lambda   8.77780
        limite lambda  14.08280
      Variable    3
        limite lambda   0.59872
               lambda   1.03672
        limite lambda   1.52372
```

(3) De nouveaux tests de normalité de Kolmogorov-Smirnov montrent que la transformation de Box-Cox a réussi à normaliser convenablement la première variable, ce qui est confirmé par l'examen de l'histogramme. Pour la seconde variable, la transformation n'a pas réussi à réduire la bimodalité des données. Quant à la troisième variable, on peut vérifier que le test de K-S ainsi que l'histogramme sont identiques à ceux affichés ci-dessus, puisque aucune transformation n'a été réalisée.

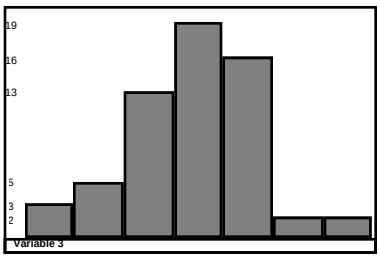
```
Tests de Kolmogorov-Smirnov-Lilliefors

 Hypothèses: r=Rejetée, Nr=non rejetée, Nc=non calculable

   Variable:    1      2      3
   Distance: 0.0700 0.2055 0.0821
   Val.Crit: 0.1144 0.1144 0.1144
 Hypothèse :     Nr      R     Nr
```



Variable 1



Variable 2



Variable 3

(4) On demande au programme de récrire les données transformées dans un nouveau fichier. VERNORM fournit les précisions suivantes sur la ligne précédée d'une flèche (⇒): les deux premières variables ont subi une transformation de Box-Cox (code a), alors que la troisième variable n'a subi aucune transformation (code 0).

```
     Transformations sur le fichier de sortie
      0=pas de transformation, a=Box-Cox, b=division en classes  c=votre choix
      de transformations, d=centrage & réduction
⇒     aa0
```

(5) On a demandé de récrire les données dans 10 espaces chacune, avec 5 décimales (format Fortran 3F10.5). Or la transformation calculée par la méthode de Box-Cox pour la seconde variable (bimodale) génère des chiffres gigantesques, qui ne peuvent pas être récrits dans 10 caractères. Le programme prend donc la liberté d'imposer le format régulier le plus économique capable

d’accommoder les données; ce format requiert 21 caractères par variable, ce qui inclut au moins un espace pour éviter que les nombres ne se touchent (format Fortran 3F21.5).

```
Champ insuffisant, nouveau format: 3f21.5
```

RÉFÉRENCES

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York. xiii + 35p.

Blanc, F., P. Chardy, A. Laurec & J.-P. Reys. 1976. Choix des métriques qualitatives en analyse d'inertie. Implication en écologie marine benthique. Mar. Biol. (Berl.) 35: 49-67.

Burgman, M. 1987. An analysis of the distribution of plants on organic outcrops in southern Western Australia using Mantel tests. Vegetatio 71: 79-86.

Cailliez, F. & J.-P. Pagès. 1976. Introduction à l'analyse des données. Société de Mathématiques appliquées et de Sciences humaines, Paris. xxii + 616 p.

Cheetham, A. H. & J. E. Hazel. 1969. Binary (presence-absence) similarity coefficients. J. Paleontol. 43: 1130-1136.

Cliff, A. D. & J. K. Ord. 1981. Spatial processes: Models and applications. Pion Ltd., London.

Clifford, H. T. & W. Stephenson. 1975. An introduction to numerical classification. Academic Press, New York. xii + 229 p.

Cooper, D. W. 1968. The significance level in multiple tests made simultaneously. Heredity 23: 614-617.

Daget, J. 1976. Les modèles mathématiques en écologie. Collection d'Écologie, No 8. Masson, Paris. viii + 172 p.

Dirichlet, G. L. 1850. Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen. Journal für die reine und angewandte Mathematik 40: 209-234.

Dow, M. M. & J. M. Cheverud. 1985. Comparison of distance matrices in studies of population structure and genetic microdifferenciation: quadratic assignment. Am. J. Phys. Anthropol. 68: 367-373.

Edgington, E. S. 1987. Randomization tests, 2nd ed. Marcel Dekker Inc., New York.

Estabrook, G. F. & D. J. Rogers. 1966. A general method of taxonomic description for a computed similarity measure. BioScience 16: 789-793.

Everitt, B. 1980. Cluster analysis, 2nd edition. Halsted Press, John Wiley & Sons, New York.

Frontier, S. 1976. Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparaison avec le modèle du bâton brisé. J. exp. mar. Biol. Ecol. 25: 67-75.

Gabriel, K. R. & R. R. Sokal. 1969. A new statistical approach to geographic variation analysis. Syst. Zool. 18: 259-278.

Galzin, R. & P. Legendre. 1987. The fish communities of a coral reef transect. Pacific Science 41: 158-165.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338.

Gower, J. C. 1982. Euclidean distance geometry. Math. Scientist 7: 1-14.

Gower, J. C. 1983. Comparing classifications. Pp. 137-155 *in*: Felsenstein, J. [ed.] Numerical taxonomy. NATO ASI Series, Vol. G 1. Springer-Verlag, Berlin. x + 644 p.

Gower, J. C. 1985. Measures of similarity, dissimilarity, and distance. Pp. 397-405 *in*: Kotz, S. & N. L. Johnson [eds.] Encyclopedia of Statistical Sciences, Vol. 5. Wiley, New York.

Gower, J. C. & P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification 3: 5-48.

Harris, C. W. & H. F. Kaiser. 1964. Oblique factor analytic solutions by orthogonal transformations. Psychometrika 29: 347-362.

Hope, A. C. A. 1968. A simplified Monte Carlo significance test procedure. J. Roy. Stat. Soc. Ser. B 30: 582-598.

Hubert, L. J. 1985. Combinatorial data analysis: association and partial association. Psychometrika 50: 449-467.

Hubert, L. J., R. G. Golledge & C. M. Constanzo. 1982. Analysis of variance procedures based on a proximity measure between subjects. Psychological Bull. 91: 424-430.

Hudon, C. & G. Lamarche. 1989. Niche segregation between American lobster *Homarus americanus* and rock crab *Cancer irroratus.* Mar. Ecol. Prog. Ser. 52: 155-168.

Isaaks, E. H. & R. M. Srivastava. 1989. An introduction to applied geostatistics. Oxford University Press, New York. xix + 561 p.

Jackson, D. A. & K. M. Somers. 1988. Are probability estimates from the permutation model of Mantel's test stable? Can. J. Zool. 67: 766-769.

Jain, A. K. & R. C. Dubes. 1988. Algorithms for clustering data. Prentice Hall, Englewood Cliffs, New Jersey. xiv + 320 p.

Kaiser, H. F. 1958. The varimax criterion for analytic rotation in factor analysis. Psychometrika 23: 187-200.

Lance, G. N. & W. T. Williams. 1966a. A generalized sorting strategy for computer classifications. Nature (Lond.) 212: 218.

Lance, G. N. & W. T. Williams. 1966b. Computer programs for hierarchical polythetic classification (“similarity analyses”). Computer Journal 9: 60-64.

Lance, G. N. & W. T. Williams. 1967. A generalized theory of classificatory sorting strategies. I. Hierarchical systems. Computer Journal 9: 373-380.

Legendre, L., M. Fréchette & P. Legendre. 1981. The contingency periodogram: A method of identifying rhythms in series of nonmetric ecological data. J. Ecol. 69: 965-979.

Legendre, L. & P. Legendre. 1984a. Ecologie numérique, 2ième édition. Tome 1: Le traitement multiple des données écologiques. Tome 2: La structure des données écologiques. Collection d'Écologie, 12 et 13. Masson, Paris et les Presses de l'Université du Québec. xv + 260 p., viii + 335 p.

Legendre, P. 1987. Constrained clustering. Pp. 289-307 *in*: P. Legendre & L. Legendre [eds.] Developments in numerical ecology. NATO ASI Series, Vol. G 14. Springer-Verlag, Berlin. xi + 585 p.

Legendre, P. & A. Chodorowski. 1977. A generalization of Jaccard's association coefficient for *Q* analysis of multi-state ecological data matrices. Ekol. Pol. 25: 297-308.

Legendre, P., S. Dallot & L. Legendre. 1985. Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. Amer. Nat. 125: 257-288.

Legendre, P. & M.-J. Fortin. 1989. Spatial pattern and ecological analysis. Vegetatio 80: 107-138.

Legendre, P. & V. Legendre. 1984b. Postglacial dispersal of freshwater fishes in the Québec peninsula. Can J. Fish. Aquat. Sci. 41: 1781-1802.

Legendre, P., N. L. Oden, R. R. Sokal, A. Vaudor & J. Kim. 1990. Approximate analysis of variance of spatially autocorrelated regional data. J. Class. 7: 53-75.

Legendre, P. & M. Troussellier. 1988. Aquatic heterotrophic bacteria: Modeling in the presence of spatial autocorrelation. Limnol. Oceanogr. 33: 1055-1067.

MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. Pp. 281-297 *in*: L. M. Le Cam & J. Neyman [eds.] Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1. University of California Press, Berkeley. xvii + 666 p.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27: 209-220.

McCune, B. & T. F. H. Allen. 1985. Will similar forest develop on similar sites? Can. J. Bot. 63: 367-376.

Mielke, P. W. 1978. Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. Biometrics 34: 277-282.

Miles, R. E. 1970. On the homogeneous planar Poisson point process. Math. Biosci. 6: 85-127.

Miller Jr., R. G. 1977. Developments in multiple comparisons. J. Amer. Stat. Ass. 72: 779-788.

Oden, N. L. 1984. Assessing the significance of spatial correlograms. Geogr. Anal. 16: 1-16.

Oden, N. L. & R. R. Sokal. 1986. Directional autocorrelation: An extension of spatial correlograms to two dimensions. Syst. Zool. 35: 608-617.

Oden, N. L. & R. R. Sokal. Investigation of 3-matrix quadratic assignment tests. (Soumis).

Orlóci, L. 1978. Multivariate analysis in vegetation research. 2nd ed. Dr. W. Junk B. V., The Hague. ix + 451 p.

Ripley, B. D. 1981. Spatial statistics. John Wiley & Sons, New York.

Rohlf, F. J., J. Kishpaugh & D. Kirk. 1971. NT-SYS. Numerical taxonomy system of multivariate statistical programs. Tech. Rep. State University of New York at Stony Brook, New York.

SAS. 1985. SAS user's guide: statistics. SAS Institute Inc., Cary, North Carolina.

Smouse, P. E., J. C. Long & R. R. Sokal. 1986. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst. Zool. 35: 627-632.

Sneath, P. H. A. 1966. A comparison of different clustering methods as applied to randomly-spaced points. Classification Society Bulletin 1: 2-18.

Sneath, P. H. A. & R. R. Sokal. 1973. Numerical taxonomy — The principles and practice of numerical classification. W. H. Freeman, San Francisco. xv + 573 p.

Sokal, R. R. 1986. Spatial data analysis and historical processes. Pp. 29-43 *in*: Diday, E. *et al.* [eds.] Data analysis and informatics, IV. Proc. Fourth Int. Symp. Data Anal. Informatics, Versailles, France, 1985. North-Holland, Amsterdam.

Sokal, R. R., I. A. Lengyel, P. A. Derish, M. C. Wooten & N. L. Oden. 1987. Spatial autocorrelation of ABO serotypes in mediaeval cemeteries as an indicator of ethnic and familial structure. J. Archaeol. Sci. 14: 615-633.

Sokal, R. R. & N. L. Oden. 1978. Spatial autocorrelation in biology. 1. Methodology. Biol. J. Linnean Soc. 10: 199-228.

Sokal, R. R. & F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon 11: 33-40.

Sokal, R. R. & F. J. Rohlf. 1981. Biometry, 2nd ed. W.H. Freeman, San Francisco. xviii + 859 p.

Sokal, R. R. & P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman, San Francisco. xvi + 359 p.

Späth, H. 1980. Cluster analysis algorithms. Ellis Horwood, Chichester.

Thiessen, A. W. 1911 — Precipitation averages for large areas. Monthly Weather Review 39: 1082-1084.

Upton, G. & B. Fingleton. 1985. Spatial data analysis by example. Vol. 1: Point pattern and quantitative data. John Wiley & Sons, Chichester. xi + 410 p.

Voronoï, G. F. 1909. Recherches sur les paralléloèdres primitifs. Journal für die reine und angewandte Mathematik 136: 67-179.

Ward, J. H. Jr. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Ass. 58: 236-244.

Watson, D. F. 1981. Computing the n-dimensional Delaunay tesselation with application to Voronoi polygones. Computer J. 24: 167-172.

Williams, W. T. & M. B. Dale. 1965. Fundamental problems in numerical taxonomy. Adv. bot. Res. 2: 35-68.