# POSEIDON

## Procédures opérationnelles en statistique et informatique pour données en langage naturel

Raymond VAN DEN DRIESSCHE*,
Ana GARCIA GOMEZ*, André GIEY**,
Anne-Marie AUBRY*

* *Banque de données pédologiques de l'ORSTOM - SSC*
*70-74, route d'Aulnay, Bondy 93140*
** *Régie informatique, Paris*

ABSTRACT

*A geographical computer system including statistical manipulations has been developed on a Univac 1108 under exec 8. Seven procedures are implemented : (1) input in plain language of 100 000 records with 500 variables of all types ; (2) retrieval and geosearch ; (3) multivariate dissimilarities between records ; (4) multivariate identification of a record ; (5) multivariate ordering of records ; (6) agglomerative clustering algorithm ; (7) Cross-tabulation with nonparametric S'test. CPU times are given for all Fortran programs.*

Un nouveau système a été élaboré au cours des deux dernières années dans les laboratoires de l'ORSTOM, l'Office de la Recherche Scientifique et Technique Outre-Mer, pour un traitement statistique par ordinateur de données géographiques. Cet ensemble de procédures opérationnelles en statistique et informatique pour données en langage naturel a reçu pour nom POSEIDON.

Le système est en démonstration à l'ORSTOM, 70, route d'Aulnay, F 93140 Bondy. sur 230 sites d'Afrique centrale. Il a, aussi, fait l'objet d'un rapport final de recherche qui détaille les procédures informatiques et statistiques, les illustre d'exemples en entrée et sortie de terminal d'ordinateur, les accompagne des temps d'unité centrale de l'Univac 1108, des programmes en Fortran, et des descriptions de sites.

*Sept procédures majeures sont opérationnelles :*

1) La caractéristique essentielle de l'*écriture sur bande* est le langage naturel. Ce sont bien les données en langage naturel décrivant chaque site qui sont perforées dans des cartes, sans ordre préférentiel, sans identificateurs, sans format, l'une à la suite de l'autre. La ponctuation disparaît.

Exemple de site perforé :

```
PROFIL/NO 6/
- HORIZON/HRZ 1//
SITE 6.8DEG S.10MIN S.24DEG E.20MIN E.SYS./1956/.SAVANE.SABLES KALAHARI.PODZOL.D
RAINAGE IMPARFAIT.ME-P17.CLIMAT AW S.1050M.22DEG.1400MM.135JOURS.
- HORIZON/HRZ 2/81361075955A//
8DEG S.35MIN S.24DEG E.7MIN E.LANDSAT-1.JUILLET 1973.SITE 6.
```

Le point sert, comme on le voit, à séparer les données. En d'autres termes, deux points délimitent la donnée et permettent à l'ordinateur de retrouver l'équivalent codé de la donnée dans un répertoire exhaustif et d'encoder la bande à l'emplacement réservé à la variable. Une suite « point espacements point » est l'équivalent d'un point ; cela facilite la correction des perforations erronées.

*Cah. ORSTOM, série Pédol., vol. XIII, n° 3/4, 1975 : 223-226.*

Les données peuvent être analytiques ou synthétiques, de terrain ou de laboratoire, collectées par des hommes ou photographiées d'avion, ou balayées par satellite. Quelle que soit leur provenance, les données sont issues de variables. La donnée est donc une des valeurs que peut prendre la variable.

Les variables sont de différents types, selon les données qu'elles regroupent : à intervalles, quand il y a une métrique sous-jacente ; ordinales, quand leurs données sont qualitatives et ordonnées ; nominales, quand, au contraire, leurs données sont qualitatives mais ne sont pas ordonnées ; binaires, quand elles sont du type absence-présence. Cela n'exclut pas la transformation d'une variable nominale en deux variables ordinales, ou, en sens inverse, l'assimilation d'une variable à intervalles peu nombreux à une variable ordinale.

Le système POSEIDON accepte 500 variables sur 10 000 sites, ou 100 000 échantillons, dans toutes ses procédures informatiques, et 400 variables sur 250 sites dans ses procédures statistiques usuelles.

Deux bandes ont été encodées pour la démonstration du système : une bande de données spatiales et une bande de données ponctuelles. Les données spatiales sont des descriptions de sites faisant appel au climat, à la végétation, au substrat, etc. et aux coordonnées en degrés et minutes ; elles sont accompagnées des identificateurs et des coordonnées 1) de l'imagerie orbitale de qualité et exempte de nuages, 2) des images radar aéroporté. Les données ponctuelles sont des profils de dix échantillons étudiés sur place et au laboratoire. Un profil est étudié au centre de chaque site. Toutes les données sont de qualité car elles proviennent de l'INEAC et de l'USGS. La perforation manuelle des 230 sites et 2 300 échantillons a nécessité 9 000 cartes. L'écriture sur bande a occupé l'unité centrale de l'ordinateur pendant 26 mn. Les répertoires ne doivent jamais dépasser 60 000 mots ; ils totalisent ici 4 000 cartes environ et furent écrits sur bande en une minute et demie.

L'actualisation d'un répertoire ou d'une bande de données, son édition ou sa traduction, ne soulèvent aucun problème, grâce aux fonctions incluses dans les deux programmes Fortran, qui occupent respectivement 68 K et 69 K de mémoire.

Exemple d'édition :

```
PROFIL/NO 385/DOSSIER 50501391B/COMMANDE 561/POUR ORTLIEB/

- HORIZON/HRZ 2/G40A04922000/FILTRE CC//
SITE 385.
27DEG N.32MIN N.
112DEG W.57MIN W.
SKYLAB-S190A.DECEMBRE 1973.
```

2) La procédure de *sélection* booléenne ajoute à l'interrogation de la bande la reproduction encodée et éditée d'un sous-ensemble qui remplit toutes les conditions imposées par l'utilisateur. Ces conditions peuvent varier ($v \leqslant 500$), mais leur nombre est limité à 100. Elles sont régies par une règle de priorité (ET prioritaire sur OU) et sont écrites en français : nom de la variable PG PP GE PE EG IG nom de la donnée. Sur programme Fortran de 80 K, l'ordinateur met un quart de seconde par site extrait.

Exemple de sélection, dans lequel 4 variables commandent 8 conditions :

```
ALTITUDE.PP.1500M.ET.ALTITUDE.IG.-1.
.ET.TEMPERATURE.GE.18DEG.ET.TEMPERATURE.PE.24DEG.
.ET.PLUVIOSITE.GE.1200MM.ET.PLUVIOSITE.PE.1800MM.
.ET.SECHERESSE.PP.60JOURS.ET.SECHERESSE.IG.-1.
```

Il est à remarquer que sans les deux conditions d'inégalité au code -1 encodé à la place des données manquantes, des sites dont l'altitude et la sécheresse n'ont pas été relevées seraient quand même extraits.

Quand la sélection porte sur une ou plusieurs variables mais ne concerne que l'aire comprise entre deux méridiens et deux parallèles quelconques, désignés en degrés et minutes, l'utilisateur est dispensé d'écrire les conditions booléennes, qui sont du reste assez complexes. Un algorithme programmé en Fortran (9 K) les écrit et les greffe sur les autres conditions.

Exemple :

```
.5DEG N.0MIN N.
.13DEG S.0MIN S.
.25DEG E.20MIN E.
.30DEG E.12MIN E.
```

Une sortie complémentaire sur l'imprimante du terminal consiste en un graphique sans distorsion portant les centres codés des images orbitales, des photographies Slar, et des sites sélectés. Ce programme occupe 17 K. Les photothèques satellite et avion peuvent ainsi être interrogées commodément.

3) Une première procédure d'analyse statistique du système POSEIDON est multivariable et non-paramétrique ; elle consiste à mesurer les $C^2_m$ *dissemblances* $d'_{ij}$ ($0 \leqslant d'_{ij} \leqslant 1$) entre les m sites ou profils ($3 \leqslant m \leqslant 250$) avec un indice nouveau,

rapide et qui ne rejette pas les enregistrements incomplets.

L'utilisateur a le libre choix des v variables ($1 \leqslant v \leqslant 400$), par cartes format, mais il a aussi à choisir, pour chaque variable, une donnée maximale. Les données sont en effet normalisées en les rapportant à leur maximum.

$$d'_{ij} = \left[\frac{1}{v} \sum_{k=1}^{v} \left(\frac{x_{ik} - x_{jk}}{x_{maxk}}\right)^2\right]^{1/2}$$

En d'autres termes, la dissemblance $d'_{ij}$ entre deux descriptions est la distance géométrique entre elles multipliée par $1/\sqrt{v}$. Si l'utilisateur choisit les maximums rencontrés dans l'aire géographique extraite par sélection, il a un programme de 17 K à sa disposition. Le programme Fortran de dissemblances, de 47 K, produit une matrice d'ordre 100 avec 400 variables en 7 mn. Les résultats sont imprimés sur le terminal. Outre la moitié inférieure de la matrice des indices de dissemblance à 4 décimales, les résultats se présentent sous une forme directement exploitable : les indices sont dans l'ordre croissant ; ils sont à deux décimales, accompagnés du nombre de variables intervenantes et surmontés du numéro de site.

4) Autre procédure statistique, *le classement* d'un site nouvellement décrit, avec v variables, dans un fichier de référence contenant m sites utilise le même algorithme que la précédente.

Les limites du programme de 15 K sont les suivantes : $1 \leqslant v \leqslant 400$ et $100 \leqslant m \leqslant 750$. C'est à un calcul et à un rangement croissant des m indices de dissemblance que procède ce programme Fortran. Les premiers sites du rangement sont ceux auxquels le site nouvellement décrit ressemble le plus. Le temps d'unité centrale est de 2 secondes pour $v = 5$ et $m = 230$ ; il est de 30 secondes pour le classement d'un profil ($v = 390$).

Cette procédure permet également d'extraire d'une bande les sites qui correspondent le mieux à un site théorique dont on a défini les caractéristiques par des données choisies ; alors que la procédure de sélection n'opère cette extraction — soit à partir de « fourchettes » (GE, PE), soit à partir de données choisies (EG, IG) — que dans la mesure où des sites identiques (et non seulement ressemblants !) existent bien sur la bande.

5) La procédure de *tri* monovariable et, même, multivariable (variables prises non pas une à une mais toutes en même temps) de m sites ($3 \leqslant m \leqslant 750$) ne diffère que peu de la procédure précédente et elle utilise les mêmes programmes. Les données maximales des v variables choisies pour le tri par cartes format ($1 \leqslant v \leqslant 400$) sont celles contenues dans la bande. Pour cette procédure de tri, il faut, en effet, calculer les m indices de dissemblance $d'_{i0}$ entre les m sites et un site fictif dont toutes les données sont mises à la valeur zéro.

Multivariable, $$d'_{i0} = \left[\frac{1}{v} \sum_{k=1}^{v} \left(\frac{x_{ik}}{x_{maxk}}\right)^2\right]^{1/2}$$

Monovariable, $$d'_{i0} = \frac{x_{ik}}{x_{maxk}}.$$

Les sites sont triés dans l'ordre croissant des indices de dissemblance. Le tri de 230 sites sur 6 variables représente 5 secondes et 7 lignes d'impression pour le programme de 15 K.

A titre d'exemple, on peut trier des profils sur la teneur en sable très grossier (fraction 1 - 2 mm) des 6 premiers échantillons (à 1, 5, 13, 25, 41 et 61 cm de profondeur) avec un maximum de 49 %. On peut aussi trier sur la pluviosité des sites.

6) La procédure statistique de *classification* de m sites ou profils, fait appel à la matrice des $C^2_m$ dissemblances multivariables entre ces sites, tels qu'ils ont été obtenus par la procédure 3.

Le regroupement des sites n'est pas systématique et myope, comme dans certains algorithmes itératifs qui se bornent, à chaque étape, à regrouper des sites sur l'indice minimal de dissemblance, pour aboutir à un seul groupe. Ici, chaque nouvelle adjonction de site à des sites déjà regroupés doit satisfaire 2 conditions : l'indice moyen de dissemblance à l'intérieur d'un groupe quelconque (de sites) doit rester inférieur, 1) à l'indice moyen de dissemblance entre deux groupes quelconques, et, 2) à tout indice de dissemblance entre un des sites regroupés et un des sites non regroupés. Le programme Fortran pour $3 \leqslant m \leqslant 100$ occupe 89 K et met 47 secondes quand $m = 100$.

7) La procédure de *tabulation* de v variables ($2 \leqslant v \leqslant 9$) en $C_v^2$ tableaux de fréquences, dont les lignes correspondent aux données d'une première variable et les colonnes aux données (maximum 34 données) d'une seconde variable trouvées sur une bande de m sites ($m \leqslant 1\,500$), est une procédure ancienne, mais toujours utile, car elle opère une réduction des données, même non-gaussiennes.

Les variables sont choisies (par cartes format) et les variables sont bornées par l'utilisateur.

Le programme de 33 K fait 10 tableaux en 2 secondes à partir d'une bande de 230 sites.

Le test $\chi^2$ d'homogénéité du tableau n'est toutefois valable que lorsqu'il est appliqué à des tableaux dont toutes les cellules renferment des fréquences dépassant 9.

L'examen attentif des tableaux peut également conduire à des conclusions intéressantes. Il est aussi loisible d'en extraire certaines lignes et colonnes, en les cumulant s'il le faut, pour en faire un nouveau tableau de fréquences, dans lequel il n'y aura plus indépendance entre les colonnes, et dont on pourra éprouver la différence éventuelle entre colonnes, par l'application du *test S'*. Ce dernier programme Fortran occupe 4 K ; il faut 5 centièmes de seconde pour un tableau 5 × 4.

## Discussion

Les coûts de fonctionnement du système sont remplacés par les temps en mn CPU d'ordinateur Univac 1108, unité de mesure plus universelle. Certains tarifs de facturation nous ont conduits à ne jamais utiliser plus de deux dérouleurs de bandes et à ne recourir à quelques pistes Fastrand que pour des occupations de courte durée.

Il a fallu se passer de coordinatographe, de table à dessiner automatique à plat ou à rouleau, d'écran cathodique conversationnel, de lecteur optique, etc., et n'utiliser que du matériel très répandu (1108 et lecteur de cartes + imprimante de terminal), afin d'assurer au système la plus large diffusion.

Le système POSEIDON est entièrement documenté. Les documents et les 13 programmes sont disponibles, à titre gracieux, sur demande individuelle des ingénieurs intéressés par ses caractéristiques ; caractéristiques qu'il faut dissocier de ses limites actuelles. Ces limites peuvent être relevées ou abaissées en fonction de la taille de la mémoire du 1108 auquel l'intéressé a accès et des objectifs de l'étude.

*Manuscrit reçu au SCD de l'ORSTOM le 25 octobre 1975,*
*Rapport résumé de la convention DRME-ORSTOM*