SAS : UN PROGICIEL A LA HAUTEUR DES BESOINS DES GEOGRAPHES ...
SAS, VERSION 5: UN TOURNANT A NE PAS RATER !

Philippe WANIEZ, Géographe.
Département E, UR 502

1. POUR UNE PLUS GRANDE MAITRISE DE L'INFORMATIQUE GEOGRAPHIQUE.

Une grande partie des travaux des géographes recourant à l'outil informatique peut être considérée, le plus souvent, comme l'application directe des méthodes d'analyse mathématiques et graphiques des matrices d'information spatiale.

1.1. Matrices d'information spatiale.

Rappelons, pour la clarté de l'exposé que l'espace objet de la recherche est divisé en unités spatiales élémentaires sur lesquelles sont relevées les valeurs (présence/absence, rang, quantité ou ratio...) d'attributs géographiques ; ils sont choisis de manière à représenter au mieux l'ensemble des situations possibles. L'élaboration d'une matrice d'information spatiale est un travail délicat, dépendant à la fois des hypothèses de la recherche et des sources accessibles. De la qualité de cette première étape dépendra grandement la valeur scientifique des résultats. La construction de cette matrice, résultant de la mise en relation de l'ensemble des unités spatiales et de celui des attributs géographiques revient à remplir un tableau rectangulaire composé de lignes figurant les unités spatiales et de colonnes représentant les attributs géographiques. A l'intersection d'une ligne et d'une colonne, on trouvera un nombre exprimant la valeur de l'attribut sur l'unité spatiale considérée (par exemple, il y a une école, la population s'élève à 1002 habitants, les électeurs sont socialistes à 63 % dans la commune numéro 03102).
Traiter ces matrices, c'est recourir à toute une batterie d'outils mathématiques et graphiques d'un usage désormais classique (I). Citons pour mémoire les techniques d'analyse des données regroupées en techniques linéaires d'une part (analyse en composantes principales, analyse factorielle des correspondances ...) et techniques de classification d'autre part. Citons aussi les méthodes de régression cherchant à rendre compte de la variation des valeurs d'un attribut par la variation de plusieurs autres, en raison de relations de causalité réelles ou postulées. Enfin, l'essentiel des techniques de la cartographie thématique permettent de visualiser les résultats des précédentes méthodes et de restituer ainsi la dimension géographique aux coordonnées factorielles, aux classes, aux résidus de régression... En résumant, un chercheur désirant procéder à l'analyse d'une matrice d'information spatiale est le plus souvent conduit à :
a - sélectionner les unités spatiales et les attributs géographiques,
b - procéder à une description statistique élémentaire des attributs

(paramètres de position, de dispersion, histogramme). Cette étape est essentielle car elle oriente parfois le choix d'une technique plus sophistiquée,
c - comprimer l'ensemble des attributs en quelques dimensions "synthétiques" (facteurs ...),
d - comprimer l'ensemble des unités spatiales en les ventilant en quelques classes significatives (au vu des facteurs, par exemple),
e - restituer les dimensions géographiques à l'aide de cartes thématiques.
Bien entendu, à ces procédés décrits depuis longtemps par Racine et Reymond (2) peuvent être adjointes d'autres techniques comme, par exemple, l'analyse de surfaces de tendances cherchant à exhiber les effets d'échelles dans la structuration de l'espace géographique.

1.2. Le nécessaire recours à l'informatique.

Il y a une dizaine d'années, la mise en oeuvre de cette méthodologie n'était pas aisée. Elle nécessitait du géographe soit des connaissances approfondies en informatique (3), soit qu'il s'en remette à la bonne volonté de l'informaticien de service, chargé de "faire passer les jobs" sans qu'il ait une bonne compréhension du problème scientifique objet de la recherche. Ces difficultés se manifestaient à chacun des stades présentés plus haut. Pour sélectionner les données, il fallait parfois écrire un programme ad hoc ; réaliser les analyses statistiques descriptives supposait utiliser un "package" particulier (le SSP IBM, très souvent) ; puis il fallait "entrer" dans une autre bibliothèque de programmes pour faire une classification ; enfin, la cartographie thématique automatique restait l'apanage des plus fortunés ou des mieux placés et bien souvent les cartes étaient réalisées à la main, réduisant ainsi l'efficience de cette méthodologie.
Depuis, les logiciels d'application ont cherché à intégrer les différentes étapes de l'analyse, supprimant ainsi les points critiques constitués par le passage d'une étape à une autre. Le Scientific Package for Social Sciences (SPSS) a constitué de ce point de vue une avancée essentielle, en proposant une méthode de gestion et de sélection des données et de nombreux programmes statistiques orientés vers l'analyse des résultats d'enquêtes sociologiques ; parmi eux, la régression et l'analyse en composantes principales ont été d'un grand secours pour les premiers géographes français usant des techniques d'analyse. Cependant, SPSS présentait deux graves lacunes : i) le système était très fermé, n'autorisant pas directement l'usage de programmes extérieurs qui auraient pu palier les carences en matière d'analyse des données par exemple ; mais surtout, point d'option graphique ou cartographique pour représenter les résultats (ce qui a été fait par la suite dans SPSSX, mais un peu tard). Le géographe devait donc continuer à jongler avec les fichiers, instructions de contrôle, programmes à compiler, assembler... et perdre ainsi du temps, de précieux et rares crédits de calculs et surtout gâcher ses bonnes dispositions vis à vis de l'outil informatique, si difficile à maîtriser.
Dans les années 1980 est apparu sur le marché français du progiciel le Statistical Analysis System (SAS) qui, en quelques mois a conquis un grand nombre de chercheurs, notamment les géographes (SAS est très utilisé à la Maison de la Géographie de Montpellier) qui peuvent maintenant maîtriser la quasi-totalité des moyens informatiques qui leur sont nécessaires. En Septembre 1985, la Version 5 de SAS ainsi que la Version 6 tournant sur micro-ordinateur IBM PC sortiront en France. Il

s'agit là d'un tournant, à bien des égards, qui va donner aux géographes des moyens accrus, particulièrement en informatique graphique. C'est pourquoi il nous a semblé utile de présenter dans une première partie les principaux atouts du système, dans sa version 82.4, en insistant sur les points les plus attrayants pour la recherche géographique. Les lecteurs connaissant déjà SAS pourront aller directement à la suite faisant le point des nouveautés de la Version 5. Il s'agit ici du point de vue d'un utilisateur et non de la simple énumération des possibilités du système qu'au demeurant on pourra trouver dans les brochures commerciales diffusées par SAS Institute S.A. (4).

2. LES PRINCIPAUX ATOUTS DE SAS.

Récemment, on a pu entendre une cartographe professionnelle déclarer que SAS n'était pas un logiciel de cartographie automatique. En effet, c'est bien plus de celà !
C'est un système de gestion d'une base de données permettant i) la saisie ou la lecture des données sur un support informatique quelconque, ii) l'extraction de cette information en vue de transformations et de traitements ultérieurs, iii) la diffusion des données brutes ou de résultats sous forme de tableaux parfois très élaborés.
C'est aussi un progiciel d'analyse statistique accédant directement aux bases de données précédentes. Statistique descriptive, bien sûr, mais aussi méthodes de régression, analyses factorielles, classifications, analyses de la variance...
C'est de plus un progiciel graphique proposant du simple diagramme à bâtons ou du camembert à la carte choroplèthe et aux surfaces tridimensionnelles.
C'est enfin un langage de programmation permettant i) d'assembler les fonctions ci-dessus en les paramétrant (les rendant ainsi d'un usage encore plus aisé), ii) d'effectuer toutes les opérations matricielles courantes, iii) d'adapter des programmes extérieurs (en Fortran ou PL/I) à l'ensemble du système.
SAS a été initialement conçu pour les ordinateurs IBM des séries 370, 30XX et 43XX sous les systèmes d'exploitation OS, TSO, CMS, DOS/VSE, SSX et ICCP ; il a par la suite été étendu à Digital Equipment Corporation sur VAX II/7XX sous VMS ainsi qu'à Data Général ECLIPSE MV sous AOS/VS et Prime PRIME 50 sous PRIMOS.
L'utilisation de ce système présente une originalité très attrayante : il s'exécute indifféremment (il s'agit des mêmes instructions) en mode interactif à partir d'un terminal écran-clavier ou en mode traitement par lots (d'un travail préalablement défini) à partir d'un fichier.

2.1. SAS gestionnaire de données.

Une base de données SAS est un fichier magnétique sans format sur disque ou sur bande. Elle est constituée d'un répertoir dans lequel sont enregistrées les caractéristiques des tableaux la composant, notamment leur nom, leur taille et leur adresse. Chaque tableau comprend un enregistrement décrivant son contenu, en particulier le nombre et le nom des variables, leurs caractéristiques (numérique ou non, longueur en caractères ...), leur contenu ; cet enregistrement descriptif est suivi des données enregistrées séquentiellement.
Une matrice d'information spatiale correspond très précisément à un tableau SAS. Chaque attribut géographique est une variable ; chaque unité spatiale est une observation. Dans le cas d'espaces emboités

(commune, canton, département, région ...), on aura intérêt à coder ces niveaux par des variables supplémentaires pour pouvoir agréger les unités appartenant à un même niveau supérieur (par addition). De même, deux variables supplémentaires pourront décrire les coordonnées géographiques des chefs-lieux (en radians de préférence) ; il sera ensuite possible de changer le système de projection cartographique. Les tableaux composant une base de données seront composés de préférence de données brutes, si possible à l'exclusion de tous les ratios, indices... pouvant être calculés par la suite.

Les supports d'information extérieurs sont très divers : il peut s'agir i) d'une bande magnétique livrée par un organisme de recensement, ii) d'un document optique (formulaire d'enquête...) directement saisi dans la base à l'aide de la procédure FSEDIT, iii) d'une base SAS ou d'un fichier SPSS ou BMDP venant d'un autre centre de calcul. Une grande variété de formats d'enregistrement sont possible allant du format caractère aux divers formats binaires en passant par les formats compactés. Quelques astuces permettent de lire et de mettre à plat des fichiers hiérarchisés comme ceux des recensements.

Outre la lecture des données, l'étape DATA autorise toutes sortes de manipulations telles que la sélection de variables par exclusion ou par inclusion explicite ; la sélection d'observations se fait très simplement à partir d'une ou plusieurs conditions. Un petit programme semble ici plus explicite :

```
DATA SELECT ; SET BASE.DONNEES ; KEEP REGION POP82 POP75 ;
                IF POP75 GT 10000 ;
```

Ce qui signifie : dans le tableau temporaire SELECT, on verse le contenu du tableau permanent DONNEES en ne conservant que les variables REGION, POP82 et POP75, et en ne retenant que les observation ayant une valeur de POP75 supérieure à 10000. De manière semblable, il est aussi aisé de constituer un tableau temporaire à partir de plusieurs tableaux permanents (ou temporaires eux aussi) :

```
DATA SELECT ; MERGE BASE.CSP BASE.ACTIVITES ; BY REGION ;
                KEEP OUVRIERS CADRES BATMENT AUTOMOBI REGION ;
                     IF REGION EQ II ;
```

Le tableau SELECT sera constitué par l'association, selon le code REGION des tableaux permanents CSP et ACTIVITE ; on ne conservera que 5 variables pour la région de code égal à II (région parisienne dans la terminologie INSEE).

Dans le cadre de l'étape DATA, de nombreuses transformations des données sont aussi possibles, soit à l'aide de fonctions (logarithme, exponentielle...), soit à partir de calculs simples s'appuyant sur les quatre opérations.

Le géographe travaillant à diverses échelles aura la faculté d'agréger les observations selon une variable de niveau, de manière extrêmement simple :

```
PROC MEANS DATA=BASE.CSP SUM ; BY REGION ;
          OUTPUT OUT=BASE.CSPAG SUM=CSPI-CSP5 ;
```

L'exécution de la procédure MEANS sur le tableau CSP conduit à la sommation de toutes les valeurs des variables CSPI à CSP5, pour toutes les observations, selon le code REGION ; le résultat est recopié dans le tableau permanent CSPAG. Cette méthode est très efficiente pour traiter des recensements fournis au niveau communal alors que le niveau d'analyse est constitué par les unités urbaines dont on connait la composition communale.

Enfin, l'impression des données sur papier est extrêmement simple. Une procédure élémentaire assure cette fonction :

PROC PRINT DATA=SELECT ;
C'est le ta bleau temporaire SELECT qui sera imprimé ou affiché à l'écran du terminal.
Notons enfin que les procédures FREQ et TABULATE permettent de calculer et d'imprimer des tableaux croisés ; leur usage est très fréquent pour le traitement d'enquêtes sociologiques.
La base de données en cours de réalisation sein de la jeune équipe CNRS P.A.R.I.S. (11) est un excellent exemple d'usage intensif de SAS. Les unités géographiques sont les agglomérations françaises de plus de 10000 habitants. Les données sont organisées en plusieurs bases physiques (plusieurs fichiers sur disque) en fonction de leurs origines (INSEE, UNEDIC, enquêtes ou publications...). Les tableaux thématiques sont tous organisés de la même manière. Toutes ces données font l'objet d'examens critiques et d'analyses approfondies sur des questions d'une brûlante actualité (par exemple : inégale qualité de la vie, disparités des coûts du logement, inégale qualification du travail...). Les résultats de ces travaux à caractère scientifique viendront alimenter une banque de données accessible à un large public (12).

2.2. SAS pour l'analyse statistique.

La description des procédures statistiques de SAS représentant plus d'un millier de pages, il est exclu d'en faire ici une présentation même succinte. Nous nous limiterons à celles qui semblent le plus utiles aux géographes.
Deux procédures assurent le calcul des paramètres des distributions statistiques ; il s'agit de MEANS et de UNIVARIATE ; cette dernière donne des résultats plus complets, mais elle est plus gourmande en ressources.
PROC MEANS DATA=BASE.POP MIN MAX MEAN STD ; VAR DEN82 ;
signifie que l'on désire connaitre, pour la variable DEN82 du tableau permanent POP, le minimum et maximum, la moyenne et l'écart-type.
La procédure STANDARD permet de centrer et réduire les variables, ce qui est souvent indispensable en cartographie thématique :
PROC STANDARD DATA=BASE.POP OUT=POPS STD=I MEAN=O ; VAR DEN82 ;
Ici, la variable DEN82 du tableau permanent POP sera centrée et réduite, puis stockée dans le tableau temporaire POPS.
Les procédures CORR et REG assurent les études de corrélation et de régression. CORR calcule et imprime les matrices de coefficients de corrélation de PEARSON ou SPEARMAN des variables données en liste.
PROC CORR DATA=BASE.POP ; VAR DEN82 DEN75 ;
Lors de l'exécution de ces instructions, on calculera la corrélation entre les densités de 1975 et 1982. L'équation de régression peut être simplement otenue de la manière suivante :
PROC REG DATA=BASE.POP ; MODEL DEN82=DEN75 ;
OUTPUT OUT=RESIDUS R=RES8275 ;
Avec ce petit programme, les résidus mesurant les principales transformations des densités entre les deux dates seront stockés dans le tableau temporaire RESIDUS sous le nom RES8275;
L'étude de la corrélation peut être avantageusement complétée par le tracé des graphiques croisant les variables endogène et exogène. Ceci s'écrit de la manière la plus simple :
PROC PLOT DATA=BASE.POP ; PLOT DEN82=DEN75 ;
Les géographes seront sûrement sensibles au fait que SAS permette de réaliser des analyses de surfaces de tendances (13). La procédure GLM admettant la formulation de modèles polynoniaux, l'estimation des

paramètres des surfaces successives jusqu'à l'ordre 4 ne pose pas de problème particulier. De nombreux tests complètent ces estimations afin de pouvoir apprécier le rôle de chacune des directions.

2.3. L'analyse des données à la française.

L'analyse statistique multidimensionnelle est très solidement représentée par les procédures PRINCOMP et FACTOR pour l'analyse factorielle, et CLUSTER pour la classification automatique. Elles sont dotées de tous les raffinements de la statistique anglo-saxonne (rotations, tests d'hypothèses ...) mais correspondent assez mal au point de vue adopté par l'analyse des données "à la française", avant tout descriptive. Ainsi, point d'analyse des correspondances, point de classification ascendante hiérarchique avec métrique du CHI-DEUX. Heureusement, SAS admet que ses utilisateurs introduisent leurs propres procédures dans le système. C'est ainsi que le Département de Mathématiques Appliquées du Centre d'Etudes Sociologiques (6) a adapté la bibliothèque de programme de l'Association pour le Développement et la Diffusion de l'Analyse des Données (ADDAD) (7). D'une richesse considérable, cette bibliothèque recule les limites de l'analyse des données tant au plan des méthodes linéaires que sur celui des techniques de classification. Tous les programmes sont utilisables sans avoir à sortir de SAS.
Pour montrer la simplicité d'utilisation d'ADDAD sous SAS, un court exemple suffit : il s'agit ici d'enchaîner quatre étapes de traitement des données, i) la transformation en rangs des variables, ii) l'analyse en composantes principales, iii) la classification ascendante hiérarchique des observations sur le plan des facteurs I et II, iiii) la partition de l'arbre de classification en 5 classes. Les unités spatiales sont constituées par 290 agglomérations françaises sur lesquelles on a relevé 8 catégories de criminalité. Voici le texte du programme SAS (8) :

```
PROC  RANK DATA=BASE.CRIM82 OUT=CRIMI82 ;Transformation en rangs puis
 VAR AUTO82PM--VVOL82PM ;                stockage du tableau résultant.

PROC ADDAD MEMBERI=ANCOMP DATA=CRIMI82   A.C.P. du tableau CRIMI82.
                          OUT=FAC ;      Les coordonnées sur les fac-
TITRE  ACP  ET CRIMINALITE ;             teurs sont stockées dans le
                                         tableau FAC.
PARAM IMPFI IMPFJ NF=2 ;                 Impression des 2 premiers fac-
                                         teurs.
GRAPHE X=1 Y=2 IP :                      Graphique des observations.
VAR   AUTO82PM--VVOL82PM  ;              Choix des variables à
                                         analyser.
IDEN CV ;                                CV est le code des villes.

PROC ADDAD MEMBERI=CAH2CO DATA=FAC       C.A.H. sur le tableau FAC.
                          OUT=HIER ;     La description de la
TITRE  CAH  SUR DEUX FACTEURS ;          hiérarchie est stockée dans
PARAM IOPT=I NPLACE=600                  le tableau HIER.
  HISTO ARBRE ;                          Impression de l'histogramme
                                         des indices de niveau de la
                                         hiérarchie.

VAR _FI _F2 ;                            C.A.H. sur 2 facteurs.
```

```
IDEN _NOM ;                              Après ANCOMP, _NOM est le code
                                         des villes.

PROC ADDAD MEMBERI=CLACAH DATA=FAC       Partition de l'arbre.
      DATACAH=HIER OUT=PART ;            Dans le tableau PART sera
TITRE PARTITION EN CINQ CLASSES ;        stockée la partition, un
                                         numéro de classe/obser.
PARAM NP=I ;                             Une seule partition demandée.
IDEN _NOM ;                              _NOM est le code des villes.
PARMCARDS ;                              La seule partition souhaitée
5                                            comprendra cinq classes.
;
```

Cet exemple ne présente qu'un squelette : plusieurs options permettent d'étudier la situation d'observations ou de variables supplémentaires. Bien d'autres techniques sont disponibles sous une forme semblable comme l'analyse déterminante, les nuées dynamiques... ainsi que des aides à l'interprétation telles que les contributions des variables aux classes d'une hiérarchie.

2.4. SAS pour la cartographie.

SAS/GRAPH est un produit complémentaire orienté vers ce qu'il est convenu d'appeler le "business graphics", c'est à dire la représentation graphique des séries statistiques. Plusieurs procédures intéressent directement le géographe :
GPROJECT permet de changer de système de projection cartographique, notamment la projection équivalente d'Albert (pour conserver les surfaces) ou la projection conforme de Lambert (pour garder les angles).
GREMOVE a pour fonction d'agréger les unités spatiales selon un code définissant le niveau d'emboitement choisi et d'effacer les contours résiduels. Associé à GREDUCE, procédure de généralisation de contours, il est aisé de réaliser une grande variété de traitements sur des fonds de carte préalablement numérisés.
La réalisation des cartes revient à trois procédures centrales dans SAS/GRAPH : GMAP, GCONTOUR et G3D.

G3D trace en trois dimensions des surfaces que GCONTOUR ne peut visualiser que dans le plan sous forme de courbes de niveau. Il est très simple d'obtenir des blocs diagrammes que l'on pourra faire pivoter autour de l'un quelconque des trois axes. Le tracé est réalisé à l'aide d'un échantillon régulier de points relevés sur lequel on connait les coordonnées géographiques (X,Y) et la valeur de la troisième dimension (Z représentant des précipitations par exemple). Un tracé standard est obtenu de la manière suivante (9) :

```
PROC G3D DATA=SELECT ; PLOT XxY=Z ;
```

Si l'échantillon n'est pas composé de points régulièrement espacés, on peut estimer les valeurs de Z sur un échantillon régulier à l'aide d'une méthode de régression. Celà se fait à l'aide de la procédure G3GRID :

```
PROC G3GRID DATA=BASE.PRECIP OUT=SELECT ;
     GRID XxY=Z/NEAR=30 NAXISI=50 NAXIS2=50 ;
```

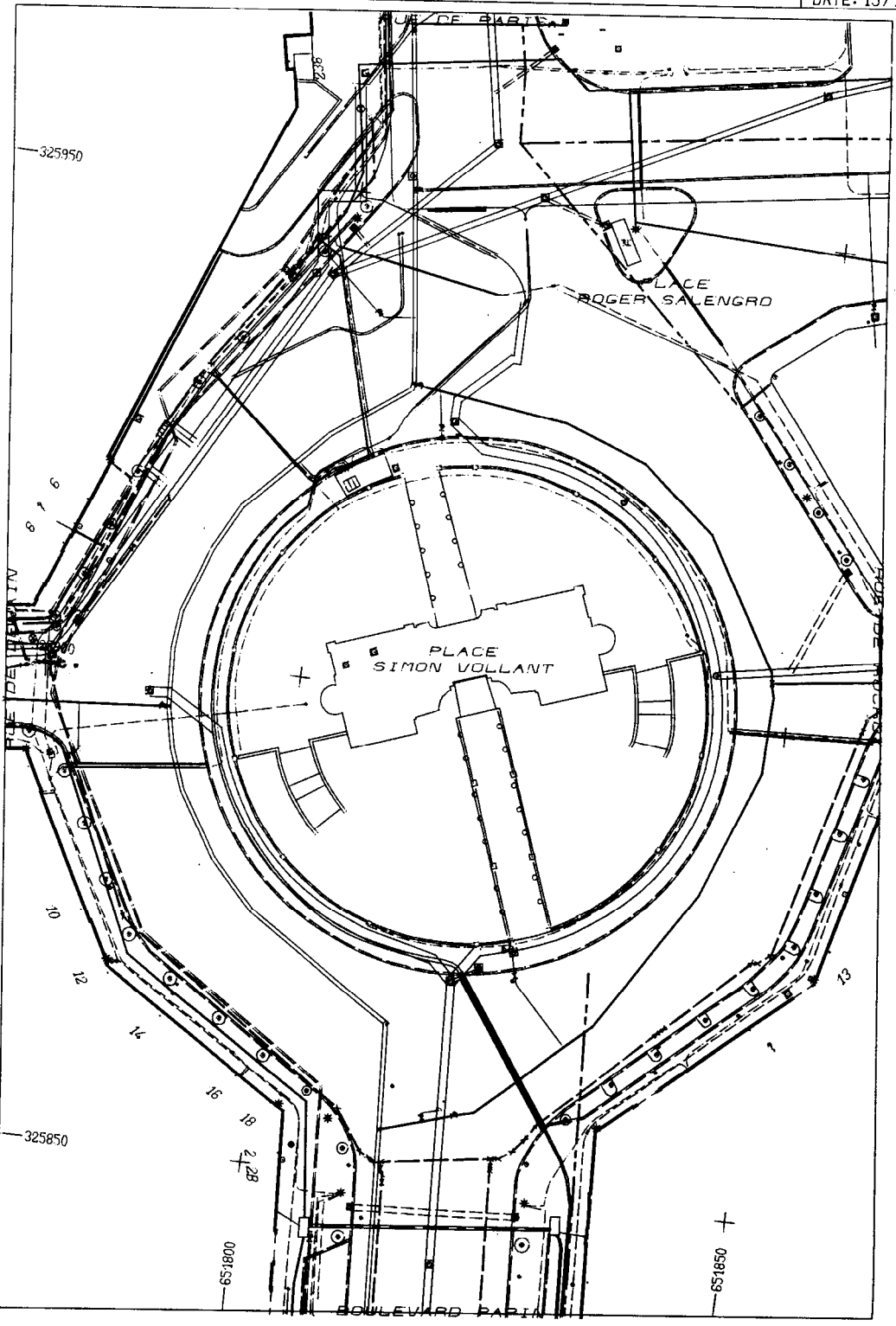
Le résultat de l'estimation sur 2500 points (50x50) sera stocké dans le tableau SELECT. La méthode d'estimation "NEAR" cherchera l'équation

d'estimation sur les 30 points d'observation les plus proches du point à estimer. La variable Z peut être obtenue à partir d'une équation de surface de tendance (par exécution préalable de GLM). Le tracé de telles surfaces devient de ce fait immédiat et particulièrement suggestif.

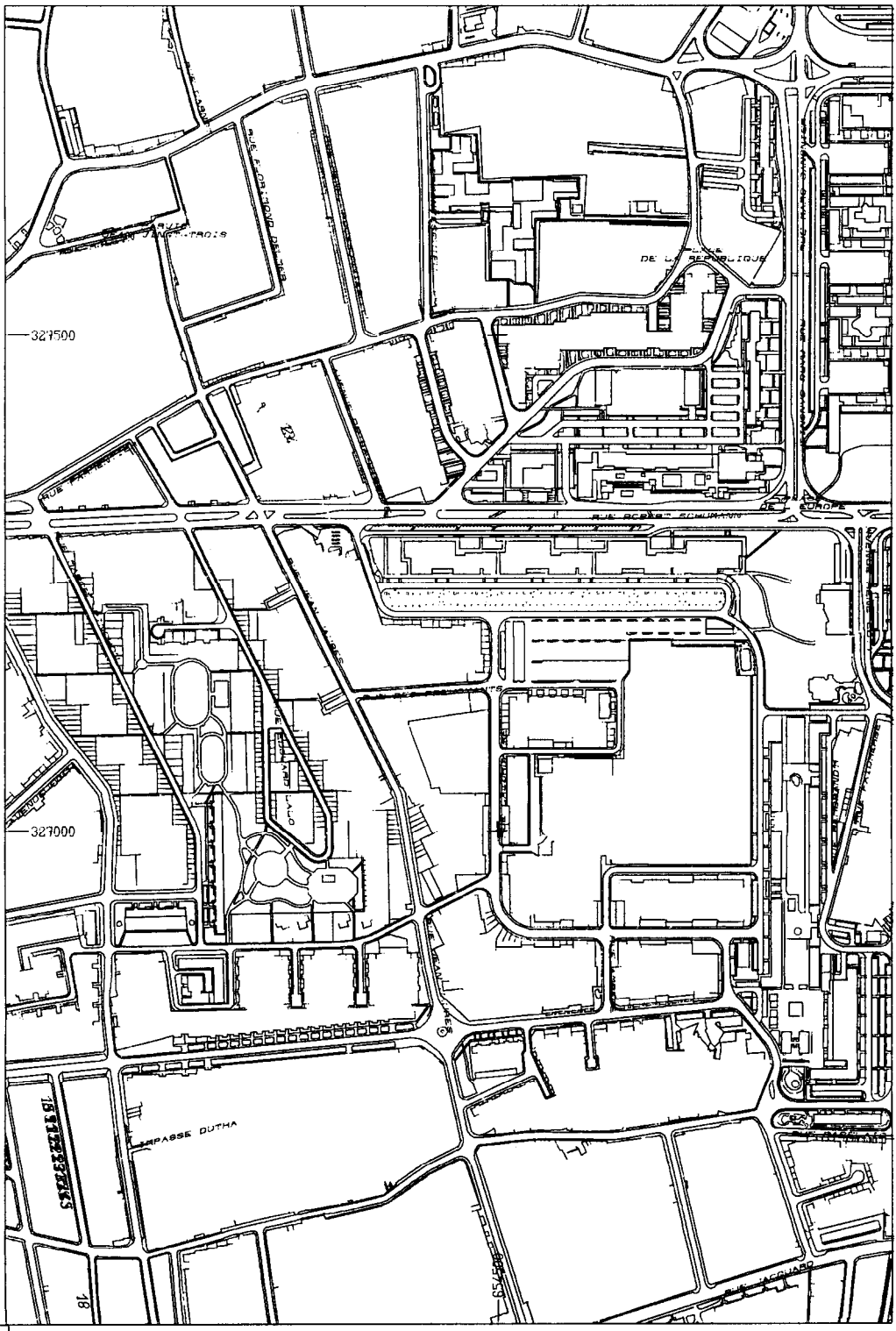
C'est probablement GMAP que le géographe aura le plus à utiliser. Elle est prévue pour tracer des cartes choroplèthes, mais quelques astuces de programmation lui font aussi tracer des cartes ponctuelles. L'appel de la procédure est extrêmement simple :
PROC GMAP DATA=BASE.POP MAP=BASE.FOND ALL ;
La ou les variables à représenter figurent dans le tableau permanent POP, le fond de carte numérisé dans FOND ; si des unités spatiales n'ont pas de valeur, le mot-clé ALL indique qu'il faudra quand même les tracer.
ID REGION ;
indique que la variables REGION, présente à la fois dans le tableau des données et dans celui du fond de carte assurera l'association entre les deux.
L'instruction de tracé comprend i) le mode de représentation (CHORO, PRISM, BLOCK ou SURFACE), ii) le nom de la variable à représenter et iii) la manière de découper en classes.
CHORO DEN75/MIDPOINTS=10 100 500 ;
ordonne le tracé d'une carte choroplèthe de la variable DEN75, découpée en trois classes dont les centres sont 10, 100 et 500. Pour découper la variable en classes, le cartographe a aussi la faculté d'employer un masque de recodage généré par la procédure FORMAT ; ceci permet de mieux maîtriser les limites des classes. On écrira alors :
CHORO DEN75/DISCRETE ;
FORMAT DEN75 FMTDEN. ;
FMTDEN est un module en bibliothèque contenant le masque de recodage préalablement défini ; il peut être temporaire ou permanent.
La définition des trames et des couleurs dépend de l'instruction PATTERN ; les trames peuvent être des hachures, des croisillons, vides ou pleins, le tout en huit couleurs.
PATTERN1 C=GREEN V=MIN45 ;
donne la définition de la trame de la première classe : elle sera verte, composée de hachures très espacées et inclinées à 45°. Il faut autant d'instructions PATTERN qu'il y a de classes à représenter.
Pour achever l'habillage de la carte, les instructions TITLE et FOOTNOTE offrent de nombreuses polices de caractères.
La procédure GMAP est d'un emploi très agréable. A partir d'un terminal graphique, le chercheur peut aisément visualiser ses résultats d'analyses multivariées (classes ou coordonnées factorielles) et faire varier les limites des classes, les figurés, les modes de représentation jusqu'à obtenir le document le plus "parlant" (celui sur lequel il a le plus de choses à dire).

L'usage de SAS/GRAPH ne dépend pas du type d'unité graphique utilisée (traceur, imprimante électrostatique, imprimante à jet d'encre, écran cathodique...). Les instructions seront les mêmes quelle que soit l'unité choisie. L'indépendance logiciel/matériel est obtenue grace aux pilotes d'unités graphiques, petits programmes adaptant le tracé virtuel en codes indiquant aux périphériques les opérations à réaliser (lever la plume, changer de couleur...). SAS fournit des pilotes pour une grande variété de périphériques graphiques. Des centres

CUDL / CDU C.D.U.	LILLE SURFACE + TOUS RESEAUX	PLACE SIMON VOLLANT	RF/0023108R/01 ECH: 1/ 500 DATE: 15/11/84



CUDL / CDU C.D.U.	MONS-EN-BAROEUL	PLAN DE SITUATION	SF/0023366X/01 ECH: 1/ 5000 DATE: 15/11/84



informatiques importants comme le CIRCE du CNRS ont réalisé des adptations pour d'autres périphériques via le logiciel graphique GPGS.

2.5. Programmer en SAS.

Nous avons présenté jusqu'ici quelques programmes n'utilisant qu'un nombre très restreint d'instructions. Le programmeur de métier désirant un jeu d'instructions plus étendu ne sera déçu ; entre le macro-langage et le langage matriciel de PROC MATRIX, il devrait trouver son bonheur ! Et si celà n'était pas, il lui resterait encore la faculté d'écrire ses propres procédures à partir d'un langage de programmation évolué comme FORTRAN ou PL/I.
Envisager ces deux dernières options nous conduirait bien loin et, à regret et faute de place, seul macro-langage sera évoqué ici. Il s'agit d'un sur-ensemble du langage SAS de base utilisé jusqu'ici. Sa principale utilité consiste en la réalisation de "macros" programmées en vue d'une application particulière et répétitive. Le chercheur peut ainsi utiliser sa propre boit à outils. Par exemple, une macro pour réaliser une carte choroplèthe (10) sera composée de deux étapes : i) la définition d'un masque de recodage et ii) le tracé à proprement parler. Nous donnons ici un court exemple d'une telle macro ;

```
%MACRO CARTO(TAB, VAR, S1, S2, L1, L2, L3, TIT) ;
DEFINITION DU MASQUE DE RECODAGE ;
PROC FORMAT ;
      VALUE FMTA LOW-&SI = &LI
                    &SI-&S2 = &L2
                        &S2-HIGH = &L3 ;

TRACE DE LA CARTE ;
PROC GMAP DATA=&TAB MAP=BASE.FOND ALL ; ID CODE ;
          CHORO &VAR/DISCRETE ;
PATTERN1 C=BLACK V=M1N45 ;
         PATTERN2 C=BLACK V=M4N45 ;
                  PATTERN3 C=BLACK V=M4N45 ;
FORMAT &VAR FMTA. ;
TITLE .F=TRIPLEX &TIT ;
%MEND CARTO ;
```

L'appel de la macro CARTO se fera de la manière suivante :

```
%CARTO(SELECT,DEN75,100,500,'10 A 100','100 A 500',
       'PLUS DE 500', 'DENSITE EN 1975) ;
```

CARTO pourra être utilisée quel que soit le tableau de données et pour toutes les variables quantitatives, pour réaliser les cartes à trois classes. A l'exécution, tous les paramètres précédés du signe & seront remplacés par leurs valeurs respectives choisies lors de l'appel.

3. UNE VERSION 5 POUR GEOGRAPHES ET UNE VERSION 6 POUR LE TERRAIN.

A la réunion de Mai du SAS European User's Group International (SEUGI), de nombreuses précisions ont été apportées sur des points essentiels du développement du progiciel. Nous avons eu, de plus la faculté d'essayer la nouvelle version (Version 5) en test au CIRCE depuis le mois d'Avril. A l'aide des informations (encore parcellaires) disponibles en cette fin du mois d'Août, nous allons tenter de dégager les nouveautés les plus saillantes, intéressant directement les géographes usagers de l'outil informatique.

3.1. Des bases d'images cartographiques.

Il est désormais possible de gérer une base d'images cartographiques à l'intérieur même d'une base de données SAS ; elles admettent maintenant plusieurs types de tableaux (autres que DATA) : les catalogues graphiques et les catalogues de "patrons". La procédure GRESPLAY, totalement remaniée, propose toutes les instructions nécessaires à la gestion de ces tableaux.
Le stockage d'une carte (comme par ailleurs de l'ensemble des images produites par SAS/GRAPH) se fait très simplement lors de l'exécution d'une procédure à l'aide du mot-clé GOUT. Par exemple, pour GMAP :
PROC GMAP DATA=SELECT MAP=BASE.FOND GOUT=BASE.CARTES ;
tous les tracés de cette procédure seront inscrits dans un catalogue graphique permanent nommé CARTE. La procédure GREPLAY assure les fonctions suivantes : (14)
i) ré-affichage de tout graphique figurant au catalogue,
ii) ré-affichage de plusieurs graphiques sur la même page à l'aide de "patrons" (templates) décrivant l'assemblage,
iii) groupement de plusieurs pages qui s'afficheront séquentiellement à la manière d'un programme de diapositives.
GREPLAY peut, bien sûr, être exécutée en mode traitement par lots, mais c'est en conversationnel qu'elle exprime toute sa puissance et sa simplicité, à partir d'un terminal de type IBM3279 graphique. A l'appel de la procédure, un premier écran s'affiche, présentant un certain nombre de champs modifiables par l'utilisateur qui peut ainsi définir : i) IGOUT : le nom du catalogue graphique contenant les images à ré-afficher, ii) GOUT : le nom du catalogue graphique qui contiendra éventuellement le résultat du ré-affichage, iii) TC : le nom du catalogue de "patrons" à utiliser ainsi que iiii) TEMPLATE : le nom du patron choisi dans TC. DEVICE : permet de préciser l'unité graphique d'affichage. Une fois ces rubriques définies, une pression sur la touche "ENTREE" provoque l'affichage du catalogue et il suffit de sélectionner les images désirées pour provoquer leur ré-affichage. Comme dans tout système interactif, les touches de fonctions (fonctions keys) sont d'une importance essentielle dans la gestion des interactions ordinateur/usager. L'appui sur la touche de fonction n°2 déclenche l'affichage du catalogue des patrons. Chaque patron est composé d'une ou plusieurs fenêtres (polygones à trois ou quatre côtés) découpant la surface d'affichage en cellules où viendront se loger les différentes images. Ce procédé est indispensable à la réalisation totalement informatisée de planches d'atlas comprenant textes, diagrammes et cartes. La saisie de la définition des fenêtres (8 au plus) se fait à l'aide d'une image sur laquelle on précise le nom du patron, le nom de la fenêtre (PANEL), sa couleur et les coordonnées en X et Y des angles du polygone exprimés en pourcentage de la surface d'affichage totale de l'unité graphique. Notons que lorsque deux fenêtres ont une partie commune, les images se superposent ; on imagine toutes les possibilités d'analyse et de démonstration de phénomènes géographiques complexes qui sont ainsi envisageables.

3.2. Un véritable langage de programmation graphique.

Avec les tableaux "ANNOTATE", SAS a mis sur pied un véritable langage de programmation graphique afin de i) adapter les sorties graphiques à des besoins particuliers, ii) programmer des applications n'existant

pas sous forme de procédure.

La première application est celle qu'il est convenu d'appeler l'habillage des cartes. Par exemple, sur une carte par points, on désire identifier chaque point. Il faudra générer un fichier contenant ces noms avec leurs localisations.
DATA SELECT; SET BASE.POP; KEEP X Y NOM;
Ici, X et Y sont les coordonnées des villes dont le nom figure dans la variable alphanumérique NOM.
DATA LOCNOM; LENGTH FUNCTION $ 8. COLOR $ 8. ; SET SELECT ;
FUNCTION='LABEL'; COLOR='CYAN'; TEXT=NOM;
Ce tableau LOCNOM sera utilisé dans GMAP de la manière suivante :
PROC GMAP DATA=BASE.POP MAP=BASE;FOND ANNOTATE=LOCNOM;
Chaque point sera identifié par son nom de couleur turquoise. Il est même possible d'écrire tout texte à n'importe quel endroit de la surface d'affichage, en plusieurs tailles et en plusieurs polices de caractères. Toutes les procédures graphique de SAS peuvent ainsi être habillées.
L'originalité des tableaux ANNOTATE réside pourtant ailleurs. Avec ce nouveau système, une grande partie des applications programmées en langage de programmation évolué, avec appel de routines graphiques (Benson, GPGS, Calcomp, GDDM...) peuvent être aujourd'hui développées en SAS. Plus besoin de s'occuper des entrées-sorties, des pilotes... Avant tout tracé, il faut définir un système d'axes en deux ou trois dimensions. Les variables XSYS, YSYS et ZSYS assurent cette fonction en plusieurs unités possibles (coordonnées absolues ou relatives exprimées à partir des données, de 0 à 100 % ou bien de MIN à MAX, ou alors à partir des dimensions de l'unité graphique). La variable FUNCTION prend des valeurs décrivant le type de tracé à réaliser, par exemple :
DRAW trace une ligne entre deux points. On peut définir la couleur, le type de ligne et son épaisseur.
POLY et POLYCONT tracent un polygone dont on peut choisir le type de ligne du contour, son épaisseur et sa couleur, la trame et la couleur de l'intérieur.
SYMBOL affiche un symbôle choisi dans une liste fixe, après avoir défini la taille et la couleur.
Chaque action fait l'objet d'une observation dans le tableau ANNOTATE ; l'exécution d'un tel programme est déclenché par n'importe quelle procédure graphique ou, à défaut, par la procédure GANNO.
Les tableaux ANNOTATE offrent donc au programmeur un outil de développement d'applications à la hauteur de ses prétentions légitimes, comme le fait déjà le macro-langage pour la construction de programmes.

3.3. Une convivialité accrue.

En commercialisant sa version 5, SAS a renforcé son image "user friendly" en développant énormément les possibilités d'interactions à partir d'un terminal "full screen" type IBM3270. Cette stratégie est clairement exprimée notamment au travers d'un nouveau produit nommé SAS/AF (Application Facility). Il s'agit d'un progiciel de construction d'applications clés-en-main où l'usager final converse avec le système central à l'aide de menus ou de pages-écrans proposant des champs variables qu'il faut remplir pour déclencher l'exécution d'une action (avec possibilité de vérification de la cohérence des réponses).
Le programmeur dispose dans sa base d'un catalogue d'écrans composés de menus, de programmes, de pages d'information... Les écrans MENU ont

pour fonction de proposer les applications disponibles et, après choix de l'usager, de provoquer l'exécution de l'application choisie.
Les écrans PROGRAM sont composés d'une part d'un écran à champs variables posant les questions nécessaires à l'exécution du programme figurant d'autre part sous forme de macros. Ces questions peuvent être par exemple "Nom du tableau à analyser ?", "Combien de classes ?"... Les valeurs transmises seront affectées aux macros variables pour exécution d'une carte choroplèthe si c'est cette option qui a été choisie au menu.
SAS/AF est une extension intéressante car tout utilisateur potentiel ne connaissant pas SAS peut quand même s'en servir et son application a été préalablement réalisée sous cette forme. C'est dans un complément très utile du macro-langage, permettant à un personnel peu qualifié d'assurer des tâches répétitives et spécifiques (la saisie des données par exemple). Pour le chercheur, celà peut être aussi un moyen de communication de ses résultats complémentaires à la rédaction d'un article dans une publication en papier ; en matière de cartographie, SAS/GRAPH mis en oeuvre conjointement à SAS/AF peut conduire à la réalisation d'atlas totalement informatisés, textes et images, faciles à mettre à jour et attrayants à consulter (15).

3.4. La version 6 ou SAS pour le terrain.

La version 5 est adaptée à des centres informatiques disposant d'ordinateurs puissants, sous des systèmes d'exploitation performants. Dans des pays développés, la connexion de terminaux éloignés sur des serveurs centraux comme le CIRCE ou le CNUSC ne pose pas de problème particulier. On utilise pour celà le réseau TRANSPAC ou des lignes téléphoniques directes à fort débit. En l'absence de tels réseaux spécialisés, le micro-ordinateur est probablement l'outil le mieux indiqué pour assurer au chercheur une puissance de calcul suffisante pour procéder aux premiers traitements. Depuis deux ans, le micro-ordinateur IBM s'est octroyé une importante part du marché, notamment en assurant la continuité matérielle et logicielle du micro vers les plus gros ordinateurs, offrant ainsi une souplesse d'utilisation inégalée. SAS Institute s'est adapté à cette révolution en présentant successivement trois produits pour le micro-ordinateur IBM.
Le premier est un programme transformant le PC en terminal graphique pouvant afficher les sorties des procédures SAS/GRAPH avec une résolution de 640x200 (avec une seule couleur) ou 320x200 (avec 3 couleurs). La transmission vers le site central se fait en mode assynchrone. Cette extension n'est pas très intéressante dans le cas de figure nous intéressant ici.
Par la suite est apparue une solution basée sur le téléchargement de SAS sur IBM PC AT/370 ou XT/370 à partir d'un système central tournant sous VM/CMS. Un menu permet de choisir la partie désirée du progiciel qui peut aussi être stockée sur diskette. L'intérêt de ce procédé réside dans la mise au point de programmes complexes sur micro (libérant ainsi le système central) puis à l'exécution en vraie grandeur sous VM/CMS. Cette solution nécessite cependant encore, mais dans une moindre mesure, la présence d'une ligne de télécommunication entre les deux machines.
En réalisant sa version 6, SAS Institute rend possible l'usage du progiciel sur micro ordinateur IBM PC XT ou AT (ou compatibles) doté de 512 K octets de mémoire centrale, d'un disque dur de 10 méga-octets et de la version 2.0 du PC DOS. On dispose ainsi d'un système de gestion

de l'écran et du display manager facilitant le dialogue, de l'ensemble de l'étape DATA et de plusieurs procédures statistiques parmi les plus fréquemment usitées. A celà s'ajoute l'Interactive Matrix Language, dérivé de PROC MATRIX et autorisant toutes les opérations matricielles, à la manière d'APL.
La version 6 semble un produit bien adapté au travail de terrain ; le chercheur peut développer lui-même ses programmes ou utiliser ceux d'autres utilisateurs. Pour l'heure, SAS/GRAPH ne tourne pas encore sous PC DOS ; souhaitons que celà arrive très vite. L'infographie nécessite cependant des moyens matériels coûteux qu'il semble assez difficile de disséminer ; en cette matière, un site central, bien équipé et doté de personnel spécialisé semble être le meilleur garant du succès.

4. LES CONDITIONS DE LA REUSSITE.

L'exposé qui précède a tenté de dégager les aspects du Statistical Analysis System les plus directement utiles au géographe ; le lecteur ne doit cependant pas y voir une forme quelconque d'impérialisme : la voie est très ouverte à toutes les formes de recherche méthodologique. faut-il, pour autant, s'évertuer à ré-inventer cet outil si puissant ? Plus utile serait la formation des chercheurs à l'usage efficient des moyens informatiques et à l'ensemble des techniques mathématiques qui les régissent. Reste donc posée la question de la formation continue, prolongement naturel de l'adjonction de ces outils modernes au corpus méthodologique du géographe, mais aussi des autres sciences sociales.

Notes et indications bibliographiques.

(1) Voir à ce propos la bibliographie de notre thèse :
WANIEZ (P.) - 1983 - Problèmes de codification et de traitement des données géographiques. Paris, Université Paris IV, 363 p.

(2) RACINE (J.B.), REYMOND (H.) - 1973 - L'analyse quantitative en géographie. Paris, P.U.F., Collection SUP - Le géographe, 311 p.

(3) La thèse de G. LEMAY en est un excellent exemple :
LEMAY (G.) - 1975 - Méthodes d'analyse chrono-spatiale : les villes de la Champagne et de la Picardie. Reims, Institut de Géographie, 271 p. + documents annexes et programmes en Fortran.

(4) SAS INSTITUTE SA, 50 Av. Daumesnil, 75012 Paris, tél. : 342-54-63.

(5) Sur les effets pervers de l'agrégation des unités spatiales, voir notre communication :
LE GAUFFEY (Y.), WANIEZ (P.) - 1983 - Stabilité et validité des facteurs sur des données géographiques agrégées. Versailles, INRIA, Actes des troisièmes journées internationales analyse des données et informatique, Tome I.

(6) Centre d'Etudes Sociologiques : 82, Rue Cardinet, 75017 Paris, tél. : 267-07-60. Contacts : M.O. LEBEAUX, I. FOURNIER et P.O. FLAVIGNY au Département de mathématiques appliquées.

(7) Association pour la Diffusion et le Développement de l'Analyse des Données (ADDAD) : 4, Place Jussieu, 75005 Paris.

(8) Ce programme est extrait de notre communication :
LE GAUFFEY (Y.), WANIEZ (P.) - 1985 - ADDAD's Library, An exemple in geography. Cologne (RFA), Réunion du SAS European User's Group International (SEUGI), 25 p.

(9) Une présentation plus complète figure dans :
LE GAUFFEY (Y.), WANIEZ (P.) - 1984 - SAS et l'analyse de surfaces de tendances. Paris, Actes de la seconde réunion du club francophone SAS, 13 p.

(10) Consulter, pour plus de précision :
LE GAUFFEY (Y.), WANIEZ (P.) - Des macros pour faire des cartes avec SAS. Nanterre, Université Paris X, Informatiques-Informations, N°6 - Juin 1984, pp. 149-174.

(11) CNRS - Jeune équipe Pour l'Avancement de la Recherche en Interaction Spatiale : 13, Rue du Four, 75006 Paris, tél. : 633-52-08. Contacts : D. PUMAIN, T. SAINT-JULIEN.

(12) Sur l'aspect particulier de la diffusion de données vers le grand public, consulter :
WANIEZ (P.) - 1985 -L'atlas télématique des grandes villes françaises. Paris, Actes du Congrès Infodial-Vidéotex, 15 p.

(13) Voir le remarquable ouvrage ci-dessous, sur l'application de l'analyse de surfaces de tendances à un pays en voie de développement : RIDDEL (J.B.) - 1970 - The spatial dynamics of modernization in Sierra Leone. Northwestern University Press, 142 p.

(14) J'utilise ici un document réalisé par H. GERARD, ingénieur au CIRCE : GERARD (H.) - 1985 - Une template dans un verre d'eau. Orsay, CNRS-CIRCE, 18 p.

(15) Un point de vue original dans :
RIMBERT (S.) - 1984 - Atlas et évolution technologique. Fribourg (Suisse) Cahiers de l'Institut de Géographie, N°2 - 1984, pp. 95-106.