LA CARTOGRAPHIE THEMATIQUE ET DES RESSOURCES NATURELLES

Marie-Madeleine Thomassin
ORSTOM, Paris
Laboratoire de Statistique du Professeur BENZECRI
Université de Paris 6, Paris

CARTOGRAPHIE AUTOMATIQUE FONDEE SUR L'ANALYSE DES DONNEES

Une application sur micro-ordinateur du programme Carthag à la régionalisation de la consommation de produits pétroliers en France (1972 - 1981).

Fonction du programme Carthag.

Ce programme est utilisé sur Macintosh, en l'occurence, mais peut être adapté avec quelques modifications minimes sur d'autres micro-ordinateurs. A partir d'un tableau de données croisant un ensemble d'unités territoriales e.g. les départements français et un ensemble de variables e.g. les consommations mensuelles d'un produit, par exemple le gazole, ce programme construit une carte, où chaque département reçoit une trame fonction du profil de la ligne afférente à ce département dans le tableau des données. Une opposition entre été et autres saisons étant apparue à l'analyse factorielle, on a choisi comme mode d'expression graphique, de symboliser par des hachurés de tonalités sombres les départements se caractérisant par une très forte consommation hivernale et très faible consommation estivale et au contraire par des hachurés de tonalités claires les départements où la consommation du gazole prédomine pendant l'été ou tout au moins, où elle est bien moins réduite qu'il n'est le cas en moyenne. Pour remplir une telle fonction et produire une carte faisant la synthèse d'informations multidimensionnelles, le programme Carthag présenté ici, doit utiliser l'ensemble des ressources de l'Analyse des Données : l'analyse factorielle des correspondances qui assigne d'après le tableau initial, des coordonnées ou facteurs à chacune des unités territoriales ainsi qu'à chacune des variables, c'est-à-dire qui traduit en termes de proximité spatiale les ressemblances existant entre départements et entre variables ainsi que les affinités des unes avec les autres mais aussi permet l'édition de graphiques-plans illustrant les résultats où figurent les sigles des variables et des individus, ensuite la classification automatique-ici classification ascendante hiérarchique (C.A.H.), qui au sein de cette représentation spatiale, détermine des classes dans notre cas des classes de départements, à chacune desquelles seront affectées ultérieurement, lors du tracé de la carte une trame unique. Le programme lui-même utilise non seulement les résultats des calculs effectués par les méthodes de l'A.F.C. et de la C.A.H. mais aussi des données numériques représentant le fond de carte. Il permet de choisir sur l'écran les trames, de façon aussi souple que possible en vue de la transcription graphique des phénomènes à représenter, qu'elles doivent symboliquement traduire, ici opposition saisonnière de la consommation du gazole. Enfin, précisons que grâce aux programme d'aides à l'interprétation, ce programme permet la représentation des classes par des caissons tramés sur les graphiques-plans issus de l'AFC et sur le graphique arborescent de la CAH - caissons tramés correspondants aux différents zonages de la carte. Ces graphiques en explicitent le contenu et lui servent de légende.

Précisons enfin que le programme se déroule en mode conversationnel.

Le présent exposé comportera 2 parties, la première intéresse la chaîne de l'Analyse des Données, la seconde, le programme Carthag proprement dit. Le support de cet exposé sera la disquette (dans le cas présent désignée par le sigle QP5) que l'utilisateur doit introduire dans l'ordinateur. Le contenu peut en être schématisé ainsi : des programmes de traitement des données, des tableaux accessibles directement à l'utilisateur, grâce à l'éditeur de texte, qui permet de les manipuler, quelles que soient leur nombre de lignes ou de pages, des fichiers numériques internes destinés à faire communiquer les programmes entre eux c'est-à-dire communiquer à un programme les résultats des calculs effectués par un autre, et éventuellement être mis sous forme de fichiers texte que l'éditeur permet d'afficher à l'écran et aussi qu'un programme particulier permet d'imprimer dans le format choisi par l'utilisateur.

La chaîne d'Analyse des Données.

Nous considèrerons successivement la création du tableau des données, la méthode d'analyse factorielle des correspondances et la méthode de classification automatique.

L'entrée du tableau des données. Examinons le tableau de la consommation mensuelle du gazole en France, extrait de la thèse de M. Moussaoui, consacrée plus généralement à l'étude de la variation saisonnière et annuelle de la consommation de tous les produits pétroliers en France de 1972 à 1981. Ce tableau comporte 95 lignes - les 95 départements rangés depuis l'Ain jusqu'au Val d'Oise, dans l'ordre de leur numéro minéralogique et 12 colonnes, les mois de l'année de Janvier à Décembre. A l'intersection de la ligne Ardèche et de la colonne Juillet, on lit 236. Ceci signifie que la consommation en gazole dans ce département au cours des mois de Juillet des années 1972-1981 a été au total de $236 \times 10^2 m^3$. La première étape d'une analyse des données est l'introduction de ce tableau dans l'ordinateur. Pour cela, après avoir introduit la disquette QP5, on appelle un éditeur de texte et l'on crée un fichier-texte en tapant les données au clavier de la machine, comme on le ferait avec une machine à écrire, selon des normes que nous préciserons en montrant une copie du tableau que nous avons entré pour traiter notre exemple. Le fichier texte que l'on crée doit obligatoirement porter un nom et est conservé dans la disquette. Celui de notre exemple porte le nom de gazole. Nous l'appelons. Il s'affiche à l'écran dans un cadre où on peut lire QP5 : gazole QP5, nom de la disquette suivi, après symbole séparateur du nom du fichier qui figurera dans l'indice des fichiers inscrit sur la disquette. Il comporte un titre choisi librement par l'utilisateur, ne devant pas dépasser 80 caractères (donc tenir sur 1 ligne). Sur la ligne suivante : nombre de colonnes suivi de la liste des sigles de 4 caractères chacun, sans blanc interposé. En-dessous le sigle du département (de 1 à 4 caractères, sans blanc) suivi des consommations afférentes aux 12 mois. Les sigles doivent être aussi évocateurs que possible. L'utilisateur demandera naturellement dans quelles limites sont comptés les nombres permis pour les lignes et les colonnes d'un tableau analysé. Pour l'ordinateur Macintosh + ou SE ou Macintosh 2, le nombre de colonnes est limité à 56, le nombre de lignes peut aller jusqu'à 500 mais il faut d'autre part que le nombre total des cases, - produit du nombre de lignes par le nombre de colonnes ne dépasse pas 8000.

Les nombres inscrits dans ces cases sont nécessairement des nombres entiers positifs ou nuls et le nombre des chiffres est limité à 9. Limitation nullement contraignante car des données supérieures peuvent facilement être ramenées, après simple changement d'échelle, à être inférieures à un milliard.
L'exemple pris pour base de notre exposé est un tableau de 12 colonnes, tenant dans la largeur de l'écran et, à fortiori, dans la largeur d'une ligne de l'éditeur de texte permettant de faire glisser le tableau sur l'écran visible. Mais si l'on considère l'ensemble des 4 produits pétroliers, carburant-auto, fuel lourd, fuel domestique et gazole, nous ne disposons plus des mêmes commodités. Il est indispensable de pouvoir écrire ce qui constitue structuellement une ligne du tableau, 48 colonnes en l'occurence, sur plusieurs lignes physiques du tableau. Le programme d'A.F.C. qui consulte ce texte créé, contenant le tableau des données est conçu de telle sorte qu'il laisse à l'utilisateur la plus grande liberté dans l'inscription des informations qu'il doit fournir, mis à part la nécessité d'entrer un titre. Il doit seulement comme précédemment inscrire le nombre de colonnes, la liste de leurs sigles, puis ce qui constitue les lignes, disposées comme des séquences en tête desquelles vient le sigle de l'individu suivi des nombres entiers représentant les valeurs des variables. Liberté des allers à la ligne. Seule contrainte : laisser un blanc entre un sigle et un nombre ou entre 2 nombres successifs. Notons qu'au lieu d'utiliser les commandes d'impression que comporte l'éditeur de texte, nous recommandons de faire appel au programme "printext" qui imprime à la suite un nombre arbitrairement grand de fichiers de texte sous le format le plus commode du point de vue du calcul : format en lignes de caractères ultra-comprimés, ce qui sur Macintosh correspond à 136 caractères par ligne.

L'analyse des correspondances. Nous examinerons successivement le programme des correspondances proprement dit, puis deux programmes qui permettent de présenter directement à l'écran ou sous forme de texte imprimé, les graphiques-plans où figurent à la fois les sigles des lignes et des colonnes du tableau initial, disposés selon les proximités que l'analyse a fait apparaître entre elles.

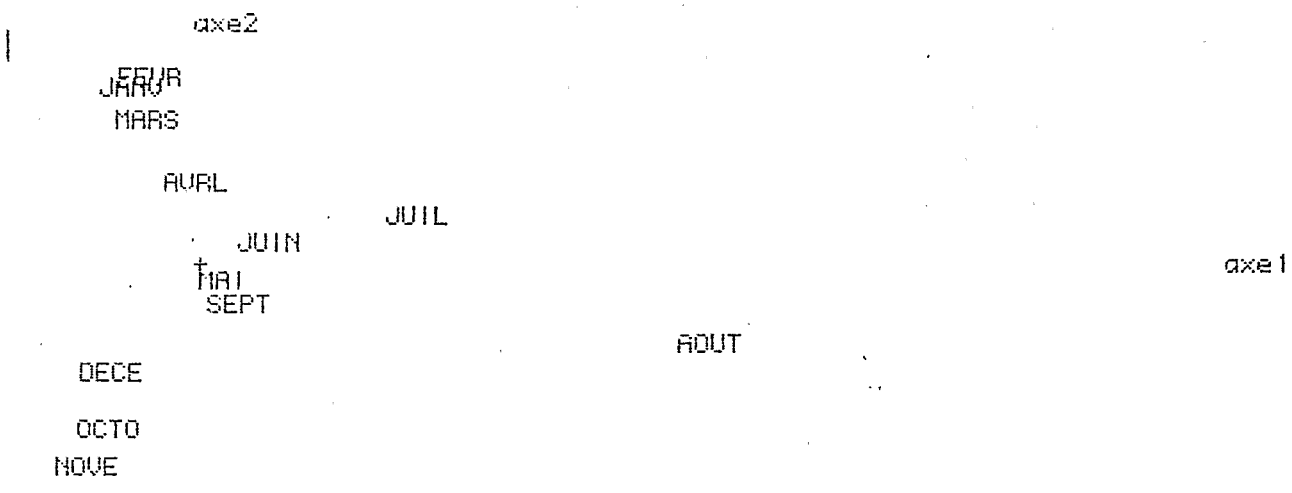
Le programme "qorils" d'analyse des correspondances. Pourquoi ce nom ? il commence par la syllabe qor - correspondances - mais les 2 lettres "qr" rappellent que les calculs de diagonalisation de matrice sont effectués par un algorithme particulièrement performant, l'algorithme SYMQR ; les lettres "il" qui suivent, indiquent que dans la présente version, le programme accepte des entiers longs (ce qui dans le jargon des informaticiens signifie que le tableau des données peut contenir des nombres allant jusqu'à un milliard et en fait même dépassant quelque peu ce nombre). Cet état de choses antérieur au programme était conçu pour utiliser uniquement le format que les informaticiens appellent "entiers" c'est-à-dire des nombres ne dépassant pas 32.767. Quant à la dernière lettre "s", elle précise que ce programme est apte à traiter des éléments supplémentaires dont nous ne dirons rien dans cet exposé. Décrivons brièvement ce programme . Après s'être informé de l'identité du fichier des données et de leur localisation (QP5 : gazole) l'ordinateur affiche le titre du tableau de données et s'enquiert du nombre de facteurs désiré pour l'analyse, 10 étant le maximum. Si l'on désire consulter les résultats de l'analyse factorielle exclusivement sur le fichier de type texte

créé par le programme (c'est-à-dire en passant par l'éditeur de texte) mieux vaut demander ce nombre si l'on s'interesse à une sortie imprimée des résultats de l'A.F.C., il faut savoir que la largeur de l'imprimante actuelle avec 136 caractères ne permet pas de dépasser l'impression des 8 premiers facteurs. Si toutefois on a dans le fichier texte des lignes plus longues, le programme "printext" imprimera le début de ces lignes où l'on pourra lire les 10 facteurs, en supprimant les fins de lignes contenant les facteurs 9 et 10. Vient ensuite le dialogue de création des éléments supplémentaires - colonnes et lignes - Pour vérification rapide, l'ordinateur peut faire défiler les 95 départements. Il indique ensuite le nombre d'itérations dont l'algorithme SYMQR a eu besoin pour calculer les facteurs, ce qui intéresse seulement les statisticiens. Il fait défiler ensuite les résultats de l'A.F.C. sous forme de lignes commençant par le sigle d'un département suivi de 10 nombres entiers affectés de signe. Ce sont les résultats de l'A.F.C. pour l'ensemble des lignes. Ces nombres représentent les coordonnées sur les axes factoriels, imprimés en millièmes. En même temps qu'il effectue cet affichage, il crée un fichier de texte "gazolecortex" que l'on pourra appeler à l'écran en utilisant l'éditeur de texte et on pourra lire tout à loisir non seulement les valeurs des facteurs mais également des informations de corrélations et de contributions, permettant de critiquer de façon précise les résultats présentés ici. A la suite des lignes relatives aux départements, il a affiché des lignes analogues relatives aux mois, de JANVier à DECEmbre. AOUT par exemple se caractérise par la valeur la plus forte sur F1 0, où il occupe une position extrême, - position traduisant le contraste des consommations de gazole entre mois de vacances et les autres mois dans les départements caractérisés par des activités touristiques intenses, alors que les activités économiques sont faibles dans la région parisienne par exemple, désertée de ses habitants en été.

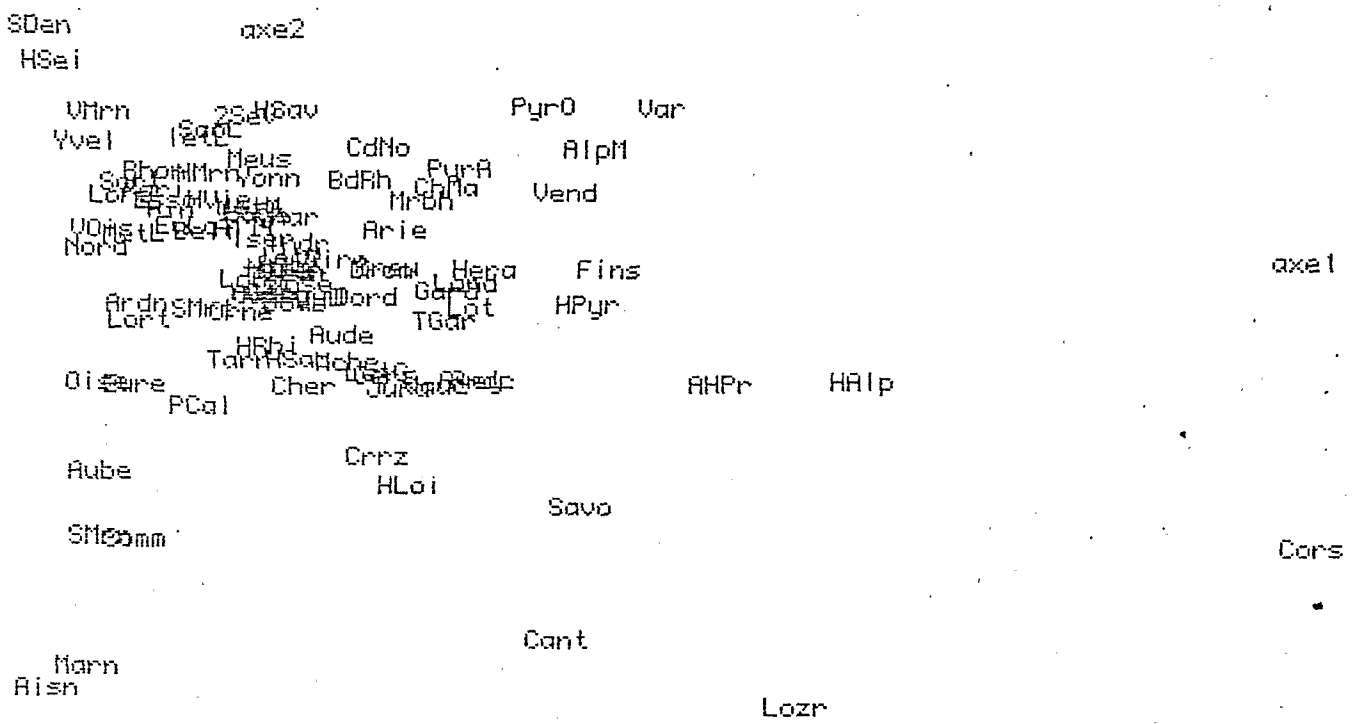
Le programme "planw" d'affichage direct des graphiques - plans issus de l'A.F.C.

Si la lecture des résultats de l'A.F.C. sur le listage du fichier "gazolecortex" créé par le programme "qorils" peut s'avérer abstraite, cette lecture sera plus facile sur des graphiques-plans créés eux-mêmes à partir de fichiers numériques, créés eux-aussi par le programme "qorils". Fichiers qui sont d'une grande utilité dans l'exécution de la classification automatique et ultérieurement de la carte proprement dite. Le programme "planw" ne crée aucun fichier mais affiche les résultats directement sur l'écran de l'ordinateur. Pour la commodité de l'utilisateur, lorsque l'ordinateur s'enquiert du fichier de base, comme pour les autres programmes la réponse à donner ne varie pas "QP5 : gazole". A partir de ce nom, par simple addition de suffixe, l'ordinateur construit d'une part les noms des fichiers numériques qu'il a à consulter pour produire les résultats demandés et d'autre part les noms des fichiers de type texte, accessibles directement à l'utilisateur et susceptibles d'être imprimés ou fichiers numériques obscurs pour l'utilisateur mais disponibles pour d'autres programmes qui produiront des résultats lisibles. "Planw" permet la localisation sur les graphiques-plans, des i (individus = les 95 départements) et des j (variables = les 12 mois) étiquetés par leurs sigles selon leur coordonnées factorielles sur les axes choisis par l'utilisateur. Toutefois, on appelle le programme

"planw" après avoir fait tourner le programme de classification automatique, le programme permet aussi l'affichage des ensembles "iq" ou ensembles des classes de départements, et "jq" ensemble des classes de mois, construits par la CAH. De là, la possibilité de représenter 1 à 4 ensembles sur les graphiques. Compte tenu des données, il est prudent de se limiter à deux et surtout de demander l'affichage des colonnes avant celui des départements car l'inverse créerait un certain encombrement de l'image, devenue très difficile à lire. Ce qui rend évidente la nécessité de la classification automatique. En effet, quant nous aurons effectué une partition de l'ensemble des départements en 8 ou 9 classes, la lecture sur l'écran des sigles de ces classes et des sigles des 12 mois se révèlera aisée. Il va sans dire que lorsqu'on crée un fichier de type texte pour représenter un graphique-plan, l'espace disponible est beaucoup plus grand et de ce fait même avec 95 départements, il y a peu de perte d'information. C'est la fonction programme "plantext", objet du paragraphe suivant de créer un fichier de type texte sur lequel s'inscrivent des informations analogues mais sous une forme telle qu'on doit y accéder par l'éditeur de texte avec cet avantage qu'on peut en demander l'impression. Lorsque l'ordinateur connaît les ensembles à représenter (n° 1 : j, n° 2 : i), il calcule ensuite les bornes de ces ensembles et pour ce faire procède à la lecture des facteurs sur i. L'intérêt de ce dialogue est de permettre à l'utilisateur de n'afficher qu'un seul des ensembles, qu'il a prévu de considérer au départ ou si 4 ensembles sont considérés de n'en afficher que deux, puis deux différents et ensuite de changer vu les répétitions permises par le programme. Si nous choisissons de considérer le plan graphique (1,2) axe 1 horizontal, axe 2 vertical, ceci signifie que les éléments de l'ensemble des colonnes (sigle des mois) et les éléments de l'ensemble des lignes (sigles des départements) s'afficheront en ayant pour abscisse le 1er facteur, calculé dans le programme "gorils" et pour ordonnées, le 2ème facteur.



Comme le montre le graphique-plan n° 1, la lecture des sigles des mois est aisée bien que FEVRier soit en surimpression par rapport à JANVier. L'opposition des mois d'été aux autres mois est clairement confirmée sur l'axe 1 tandis qu'automne et hiver s'opposent sur l'axe 2, opposition qui sera plus clairement expliquée lorsqu'on examinera le plan (2,3) et qu'on avancera dans les résultats.



Le graphique-plan n° 2 reproduit le nuage des sigles de départements, particulièrement dense au centre par contre à la périphérie une quarantaine de sigles de départements se lisent clairement. A l'extrême droite associée au mois d'août on lit CORS. La Corse non subdivisée dans cette étude en nord et sud se caractérise par une forte activité de vacances. Suivent les départements des Hautes-Alpes, des Alpes de Haute-Provence et en bas de l'écran à droite les départements de la Lozère, de la Savoie et du Cantal. Sur le bord gauche de l'écran, on trouve en haut associés à l'hiver les départements de la couronne : Seine-Saint Denis, Hauts-de-Seine, Val de Marne, Yvelines etc... Au contraire en bas, on lit Aisne, Marne, Somme, Seine et Marne, - départements qui ont une consommation en gazole particulièrement forte en Octobre, car la récolte des betteraves sucrières et leur élaboration y occupent plusieurs semaines en automne. L'examen du plan 2, 3 corrobore les résultats.

Le programme "plantx" de création des graphiques-plans sous forme de fichiers texte.

L'ordinateur pose les mêmes questions relatives au fichier de base (QP5 : gazole) et aux ensembles considérés, que l'on fixe à deux : i et j. On rentre dans le programme précédent pour le calcul des bornes des facteurs sur i afin de préparer le tracé. La représentation de j et de i ayant été demandée, voyons en quoi ce programme diffère du précédent - par le choix des dimensions du graphique = largeur fixée au maximum à 136 caractères - hauteur en lignes fixée, si l'on veut avoir même échelle sur les 2 axes à 19 lignes sur l'écran, et 14 sur le listage.

"La hauteur du graphique" ◘ s'avère une question embarassante compte tenu de ces données car il nous faut dilater largement l'échelle de l'axe 2, ou l'importance prédominante de l'axe 1 sur tous les autres axes : forte opposition de l'été aux autres mois alors que le contraste hiver-automne ne se traduit que par des distances assez faibles sur l'axe 2. L'ordinateur crée un fichier de texte contenant le plan demandé. L'utilisateur peut en demander d'autres. Notons toutefois que l'ordinateur a placé le fichier texte sur la disquette QP5 que nous utilisons et lui a donné pour titre "gazoleplantx". Le suffixe a été ajouté au nom du tableau de données sur l'éditeur de texte. L'affichage du plan (1,2) issu de "QP5 : gazoleplantx" ne montre qu'une partie de la projection du nuage sur ce plan. Commentons brièvement quelques résultats d'AFC : min et max sont les valeurs minima et maxima du 1er facteur pour l'ensemble des départements ; lam désigne la valeur propre correspondant à la lettre grecque λ, ici très faible 1,36 millième seulement parce que les contrastes entre départements sont peu prononcés même s'ils sont très significatifs, le taux est de 6,89 dixièmes ou encore 69 %. Il se confirme que l'essentiel de l'information contenue dans le tableau analysé est représenté sur l'axe 1 - suivent des informations semblables relatives à l'axe 2. Si l'ensemble des éléments de j a été représenté, une vingtaine de départements n'ont pû être placés sur le graphique parce qu'ils se seraient superposés à d'autres. L'utilisateur peut connaître de 2 façons la place de ces départements 1) en consultant le fichier texte "gazolecortex" qui en donne les coordonnées numériques et d'autre part en consultant ultérieurement les résultats de la classification automatique qui mettront ces départements dans la même classe que d'autres effectivement représentés et dont ils sont très proches et auxquels précisément ils se seraient superposés si on avait tenté de les imprimer. La présentation graphique montre combien était justifié notre demande d'un nombre important de lignes (50) et aurait dû être supérieure (70). Les sigles des mois ne se superposent ni ne se recouvrent car la création du fichier texte a été faite de telle sorte que rien de tel ne se produise, avec toutefois un inconvénient que certains départements manquent. "Planw" avec affichage direct à l'écran, opérait tout autrement. Il mettait sans vérification tous les sigles quitte à créer au centre du graphique après affichage des départements un fouillis inextricable.

Le programme "CHrbz" de classification ascendante hiérarchique.
Ce nom évoque toutes les fonctions du programme. "CH" : programme de C.A.H., "rb" rappelle que le programme dessine des arbres. A propos de la lettre "z", on signalera seulement qu'elle comporte la création d'aides à l'interprétation, très utiles pour le praticien de la classification automatique. Bien que toutes ces fonctions soient intégrées dans le programme, nous distinguerons trois paragraphes : la CAH, l'arbre, les aides à l'interprétation.

La C.A.H. L'ouverture du programme s'accompagne de la question et réponse usuelles. Comme le programme accomplit de nombreuses fonctions par exemple, modifier un tracé, calculer une partition ou des aides à l'interprétation, il s'enquiert de la nécessité de classer l'ensemble des i - des départements, ce qui peut surprendre. Le programme effectue une classification sur les facteurs de l'A.F.C. et pour ce faire, consulte un fichier créé par le programme d'analyse ("gorils"). Mais pour que "CHrbz" s'exécute, il a besoin de connaître le nombre de facteurs à utiliser.

Il les lit et par des calculs de distance range les uns avec les autres les éléments qui sont le plus proches. La CAH procède en mettant deux individus ensemble pour constituer un noeud, puis deux autres pour constituer un deuxième noeud ou encore en agrégeant au premier noeud constitué un troisième individu et ainsi de suite. Ainsi par 94 agrégations successives se constitue la hiérarchie des départements, culminant avec le noeud 94. En nous plaçant à un certain niveau dans cet arbre, nous choisissons une partition de 8 ou 9 classes, à chacune d'entre elles on affectera une trame. Le programme affiche la création des noeuds. Sur chaque ligne figure après le numéro du noeud, le nombre de maillons utilisés dans la chaîne de création et le niveau de création du noeud. Si cette information intéresse surtout les statisticiens, il importe de signaler que ce programme de CAH procède par recherche en chaînes, pour chaque individu ou chaque noeud, de ses plus proches voisins. Le niveau inscrit représente en bref, le degré de proximité des individus que l'on a pu agréger. La fin de cet affichage et le début des résultats de la CAH sont reproduits ci-dessous :

```
n= 92 carm=  2  niv= 1.80e-4
n= 93 carm=  2  niv= 3.78e-4
n= 94 carm=  1  niv= 6.52e-4
```

c	189	188	187	186	185	184	183	182	181	180	179	178	177	176	175
T	3315	1919	914	541	351	243	229	191	154	150	117	110	93	83	74
A	186	187	183	180	20	173	170	179	169	163	164	29	168	50	154
B	188	185	184	174	181	177	182	176	178	171	175	161	172	159	167

Les résultats de la CAH se présentent sous forme de blocs de 4 lignes, précédées de "c" désignant le numéro de la classe, de "T" le taux d'inertie, de A puis de B indiquant les numéros des deux descendants du noeud. Quelques commentaires sans reprendre la théorie de la CAH : "c" 189 n'est autre que le sommet, le 94ème. Si l'on range les sommets à la suite des 95 départements, on voit qu'il reçoit le n° 189 = 94 + 95. "T" désigne un taux de représentation de l'information en dix-millièmes. "3315" signifie donc qu'1/3 de l'inertie, de l'information contenue dans le tableau de données peut être schématisé par la partition des 95 départements en 2 classes ; ces 2 classes étant respectivement A et B, 186 et 188, dont il sera possible ultérieurement de considérer le contenu. Cette opposition est une opposition entre mois où les activités de vacances entraînent une consommation de gazole très importante en août et mois où cette consommation est très faible, l'activité économique étant développée en automne et en hiver. La comparaison de ce taux 1/3 environ au taux de 70 % trouvé pour le premier axe est intéressante. Il représente une opposition graduée de l'ensemble des départements, opposition que traduira la gamme de grisés choisie. Il donne plus d'information que n'en fournit une simple coupure en 2 classes.

Arbre et partition. Jusqu'à présent, le calcul effectué par le programme tout en représentant l'essentiel de classification, n'a pourtant créé que des fichiers numériques, inaccessibles à l'utilisateur. Le programme doit donc faire l'arbre et la partition sur i en la définissant soit par les noeuds les plus hauts ou des noeuds spécifiés. Il s'agit de retenir de la CAH qui comporte au-dessus des 95 éléments, 94 noeuds, une sous-hiérarchie à la base de laquelle se trouvera une partition, par exemple en 9 classes et au-dessus de ces classes des réunions successives en classes supérieures

jusqu'à parvenir au sommet. Pour procéder à un tel choix de façon conséquente, une connaissance approfondie des résultats de la CAH est nécessaire. C'est pourquoi l'on prévoit de rentrer après première consultation dans le programme "CHrbz" pour demander à bon escient une sous-hiérarchie. La partition définie par les noeuds les plus hauts, comporte les noeuds auxquels correspondent les pourcentages d'explication les plus forts. Le nombre de classes choisi, varié de 2 à 100. Compte tenu de la connaissance préalable des données et aussi pour satisfaire aux exigences de la représentation cartographique ultérieure, nous retenons 9 classes. Mais ce nombre aurait pû être déterminé par les valeurs de Taux. Le programme ensuite crée un fichier de type texte donnant l'arbre de cette partition, d'après la présentation demandée par l'utilisateur (largeur en caractères choisi pour le tracé, nécessité de passer une ligne entre 2 individus). Alors qu'il crée ce fichier, il liste les sigles de tous les départements. Utile vérification de ce qui a été fait jusqu'à présent, et propose le tracé de l'arbre de la CAH générale qui comprend 95 lignes -
La consultation de ces arbres sur le listage du fichier de texte "gazole.arbx" est très claire, (listage imprimé par le programme "printext"). Pour le tracé de l'arbre général, il convient de demander la largeur maxima permise par l'ordinateur.

Aides à l'interprétation de la classification ascendante hiérarchique. Les listages d'aide à l'interprétation sont d'une véritable utilité au cartographe. Ils lui permettent de connaître de façon fine les caractéristiques des classes de départements créés et d'autre part, c'est en calculant ces aides à l'interprétation que les fichiers numériques indispensables au programme de cartographie pour effectuer un certain nombre de constructions, sont créés. Faire Facor et Vacor sur i est donc indispensable - opération qui achève le traitement des individus. L'ensemble des traitements pour i peut être effectué par l'utilisateur en poursuivant le dialogue, de façon semblable pour j, c'est-à-dire les variables - les mois - Mais ce qui intéresse le cartographe c'est de représenter sur la carte les résultats d'une classification qui se traduiront par des zonages différenciés. Interpréter la classification et donc la carte en termes de variables, lorsqu'on a un ensemble j comprenant 12 mois, pour i, est facile. Par contre, dans le cas du tableau des consommations mensuelles des 4 produits pétroliers précédemment évoqué et comportant 48 colonnes, l'interprétation directe en termes de ces 48 variables s'avère impossible. Il faut donc procéder à l'agrégation de ces variables et considérer en quelque sorte un tableau intermédiaire donnant par exemple au lieu des consommations mensuelles pour chacun des 4 produits considérés, leur consommation pour des périodes plus longues, variant selon les produits (saison entière, moitié de l'année, trois-quarts d'année ... etc) opposées à l'été. Ainsi, placé en présence d'un tableau qui a une dizaine de colonnes et de pourcentages calculés sur ce tableau par le programme d'aide à l'interprétation et disposant de graphiques-plans, où s'affichent les sigles de ces agrégats de variables, l'utilisateur comprendra sans peine la signification, l'interprétation des classes de départements ainsi créés. "Faire Vacor sur jq pour i" en l'occurence signifie tout simplement représenter, interpréter le tableau dont les lignes sont les i, les départements mais dont les colonnes ne sont plus les mois mais les classes de mois reconnues pertinentes par la CAH.

"Faire Facor sur iq pour j" est tout à fait secondaire, la représentation de j étant claire. Mais si on voulait interpréter les classes de l'ensemble j, cette interprétation ne pourrait se faire directement en terme de départements mais en termes de groupes de départements, ce qui justifie alors la considération d'un tableau dont les colonnes sont les j eux-mêmes, les colonnes primitives du tableau de données mais dont les lignes sont réalisées par cumul de départements suivant les classes créées par la CAH.
En conclusion, le programme "CHrbz" a créé des fichiers de type texte contenant d'une part l'arbre de la CAH (gazoleiarbx et gazolejarbx dans notre exemple) et d'autre part les listages. VACOR et FACOR que nous ne commenterons pas dans cet exposé, renvoyant le lecteur à la bibliographie.

Le programme CARTHAG de cartographie automatique.

Son exécution présuppose, nous l'avons dit, l'analyse par A.F.C. et par C.A.H. d'un tableau de correspondances ou tableau de nombres positifs dont l'ensemble des lignes est l'ensemble des 95 départements et l'ensemble des colonnes, l'ensemble des consommations mensuelles du gazole. Pour faire la carte, outre ces informations statistiques, il faudra traiter des informations proprement géographiques afférentes à la numérisation des contours permettant de tracer le fond de carte. Donc, pour plan de ce paragraphe, nous présenterons la création des fichiers numériques indispensables à l'ordinateur pour ce tracé, ensuite nous considèrerons l'exécution du programme "Carthag" avec les options qu'on y a mises, enfin nous indiquerons le moyen d'utiliser "Carthag" pour représenter une partition d'unités territoriales en classes ne résultant pas de l'Analyse de Données ou obtenues sur un autre ordinateur ou par une autre méthode, sans qu'il soit possible de transférer les fichiers nécessaires à l'éxécution normale du programme Carthag, de là, la nécessité d'entrer au clavier sur l'éditeur de texte, la description de la partition.

Création des fichiers du fond de carte. Elle procède en 3 étapes : 1) le cartographe numérise la carte, de grand format de préférence (France 1/1000 000) 2) l'utilisateur entre en créant 2 fichiers de type texte les informations numériques relevées sur la carte 3) par l'exécution d'un programme "litxcarte", ces fichiers de type texte sont vérifiés et mis sous forme d'un fichier numérique "Francereg" qu'utilise le programme "Carthag".

Préparation des données numériques sur la carte. Le cartographe utilise des pastilles adhésives de 2 tailles. Les plus grandes lui serviront à étiqueter les départements et il y portera le numéro minéralogique, il collera les plus petites sur la carte aux points qu'il aura choisis pour réduire en un tracé polygonal l'ensemble des limites. A titre d'indication, une carte de France même schématique a nécessité 251 points, une carte de Grèce plus de 1000 points. Les zones figurées hors du champ de la surface cartographiée(carton à échelle agrandie de la région parisienne ou iles grecques, par exemple) sont décrites par sommets et par liste ordonnée de sommets sur les contours exactement comme on le fait pour les unités territoriales de la carte principale. Dans sa version actuelle le programme prévoit un maximum de 2000 points, maximum qui pourrait être dépassé sans grand effort pour la programmation. Le nombre d'unités territoriales est limité à 200, limitation ne présentant pas de gêne pour les applications

que l'on peut être tenté de faire à l'échelle que permet l'écran de l'ordinateur ou la largeur de l'imprimante, sur laquelle est reproduit le contenu de cet écran. Les points sont donc numérotés de façon séquentielle, par exemple dans le cas d'une carte de France schématique de 1 à 251, puis pour chacun des points, on relève ses coordonnées. Elles peuvent être relevées en utilisant une échelle quelconque (mm ou cm) pourvu que ces coordonnées soient des nombres entiers et soient comptés de la façon suivante : abscisse (x) à partir du bord gauche de la feuille, ordonnée (y) à partir du bord supérieur de la feuille. Autrement dit, les coordonnées sont positives, l'origine est placée au Nord-Ouest. Dans les fichiers que nous avons créés, les mesures sont exprimées en mm. L'utilisateur n'a pas à se préoccuper de la dimension globale de la carte qu'il utilise car le programme de création de fichiers numériques met lui-même les nombres à l'échelle de façon commode pour l'exécution du tracé de carte proprement dit.

Les fichiers texte du fond de carte. Ils sont au nombre de deux. On leur donne pour nom celui du pays dans notre cas : France, suivi respectivement des suffixes "stx" et utx". "s" signifie sommets "u" unités territoriales, "tx" précise qu'il s'agit de fichiers de texte. Ouvrons le fichier "QP5" : France stx" tel qu'il a été créé à l'éditeur de texte. Il comporte une ligne de titre choisie librement et en-dessous viennent les sommets avec leurs coordonnées.

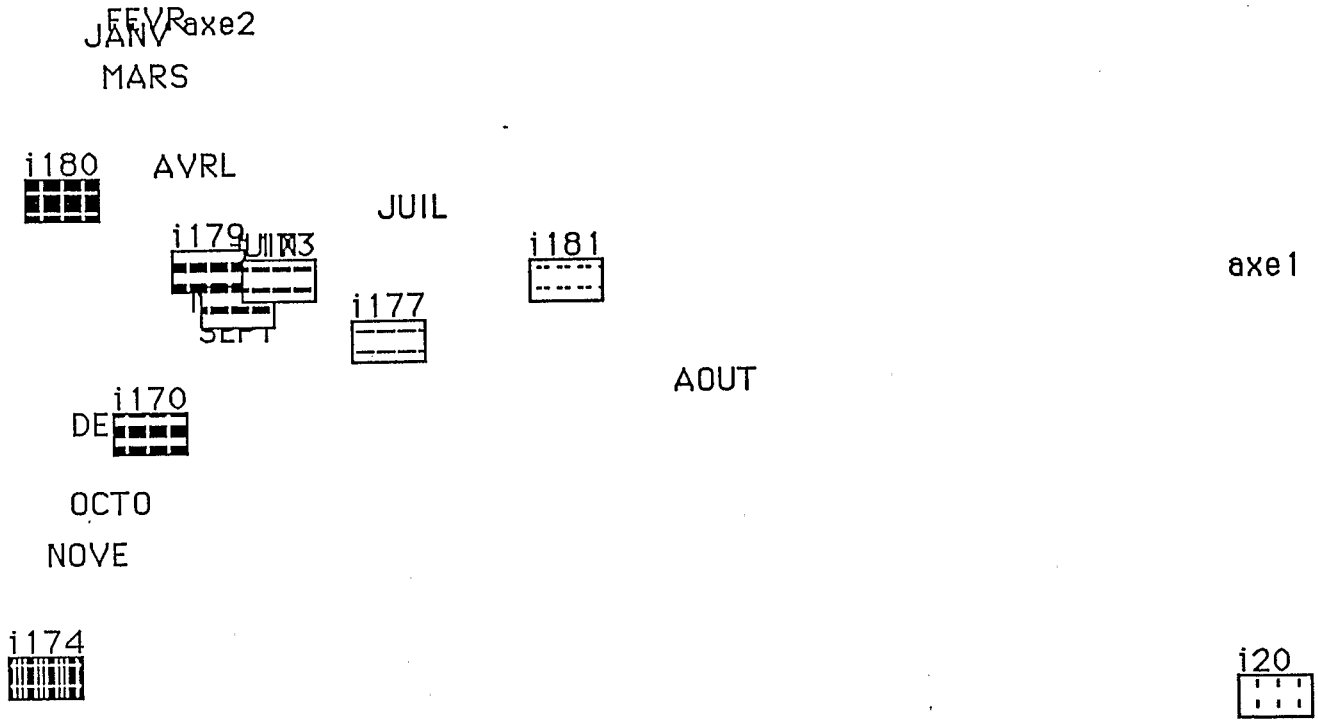
"FRANCE coordonnées des sommets placés sur les contours "
S1 225 800
S2 315 795
S3 310 870
etc... jusqu'à
S251 et deux coordonnées.

Ceci signifie que la petite pastille portant le chiffre 1 se trouve en fait placée à 225 mm du bord gauche de la carte de France utilisée et à 800 mm de son bord supérieur. Donc la numérisation a commencé par le Sud-Ouest de la France. Signalons ici les libertés laissées à l'utilisateur. Seule compte l'inscription des numéros de sommets, consécutivement, sans laisser de vide, par exemple ici de 1 à 251, avec après chaque sommet deux nombres entiers. Les allers à la ligne sont laissés à la commodité de l'utilisateur. A vrai dire la lettre "s" elle-même, figurant devant le numéro du sommet, n'est pas prise en compte par l'ordinateur qui considère seulement les chiffres et les blancs séparant ceux-ci. S'il le voulait, l'utilisateur pourrait donc supprimer la lettre "s" ou encore écrire "sommet". La forme choisie n'est donc que simple suggestion. Affichons sur l'écran le fichier "Franceutx". Là, encore un titre et à la ligne suivante on lit les données :
"FRANCE numéros des sommets sur les contours des départements"
"ul 8s 71 83 85 86 89 183 70" Ces nombres concernent l'unité territoriale, le département n° 1 dont le contour est assimilé à un polygone à 8 sommets ("8s"). Ces sommets portent respectivement les numéros énumérés de 71 à 70. Il est essentiel que l'utilisateur donne d'abord le numéro du département, puis le nombre de sommets sur le contour de ce département et ensuite les numéros de ces sommets. Ceci permet à un programme ultérieur qui crée les fichiers numériques destinés au programme de cartographie, d'effectuer toutes les vérifications nécessaires. Il est en effet souhaitable que le fichier numérique ne puisse contenir aucune erreur excepté, cela va sans dire , les erreurs sur les coordonnées que l'ordinateur n'a aucun moyen de vérifier.

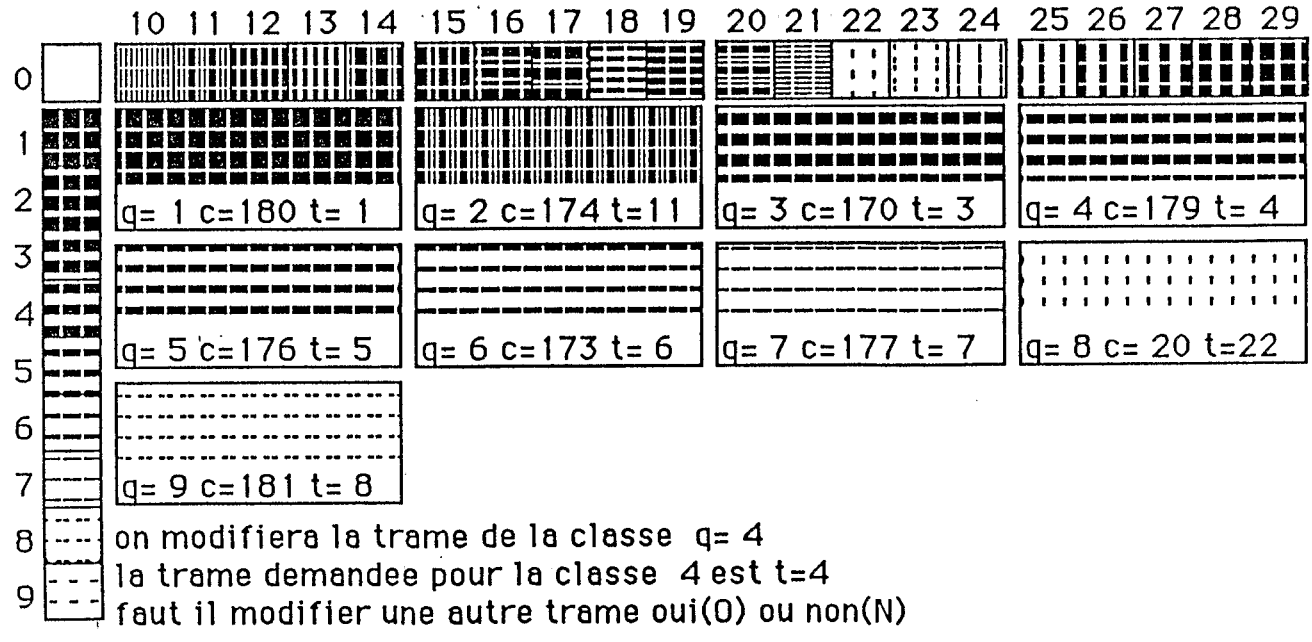
Comme pour le fichier "Francestx", les lettres que nous avons introduites : "u" pour dire en tête de ligne qu'il s'agit des informations relatives à une unité territoriale ou département, "s" après le chiffre 8, pour rappeler que c'est le nombre de sommets qui est donné et non un numéro de sommet, sont facultatives. De même, la disposition en lignes est, elle aussi, facultative, une fois le titre du tableau entré.

Le programme "Litxcarte" créant le fichier numérique du fond de carte. Le programme demande à l'utilisateur le nom du pays, le nombre des unités territoriales (95), le nombre de points relevés sur les contours (251) et affiche le titre, "FRANCE coordonnées des sommets placés sur les contours", ce qui prouve que le programme a eu accès au fichier précédemment décrit. Après que l'utilisateur ait entré un caractère quelconque, la lecture du fichier des coordonnées de sommets s'effectue sans aléa, à condition toutefois que le contenu de ce fichier ne présente pas de contradiction interne et que les nombres se présentent sous la forme "s, x, y" sur "s" de 1 à 95 et sous forme de coordonnées affectés à ces nombres. S'il en est autrement (ligne sautée, omission de coordonnées) le programme de lecture s'en aperçoit et envoie 1 message d'erreur à l'utilisateur en vue de la correction du fichier de texte.

Le programme "Carthag" de cartographie automatique. L'ouverture de ce programme s'accompagne à l'écran d'une série de questions ayant trait au disque utilisé (QP5) au nom du pays, (FRANCE), au nom du fichier de données (gazole), à la nécessité de présenter les résultats de l'AFC. En répondant par l'affirmative, nous voyons s'exécuter à l'écran le programme "planw", décrit précédemment, mais ici le nombre des ensembles considérés va de 1 à 5. Voici pourquoi : outre les ensembles "i", départements, "j" : variables, "iq" classe de départements, "jq" classes de variables, est prévue "iQ" classes d'unités territoriales avec trame. Les numéros des classes de départements précédés de la lettre i, comme il81 (-classe n° 181 de départements dont le contenu figure sur le listage de CAH ("gazoleiarbx") ; et apparaîtra sur la carte, s'affichent au-dessus d'un caisson dont la trame attribuée caractérisera les départements de cette classe. Si l'on demande l'affichage pour les axes 1,2, des 3 ensembles j, iq et iQ, on obtiendra successivement : d'abord le plan-graphique des mois avec ensuite superposition des numéros de classe et enfin des caissons tramés. La lecture de ce plan, reproduit ci-dessous peut se révéler parfois difficile compte tenu des superpositions successives.
Examinons le plan (1,2) et les suggestions de trames qu'il contient. Les classes associées à l'été i20, il81, il77 doivent être affectées de trames claires dont le grisé croît au fur et à mesure que l'activité estivale décroît. Les classes il74, il70, il79, il80 sont peu intéressées par les activités de vacances, leurs activités étant liées à l'hiver et à l'automne.

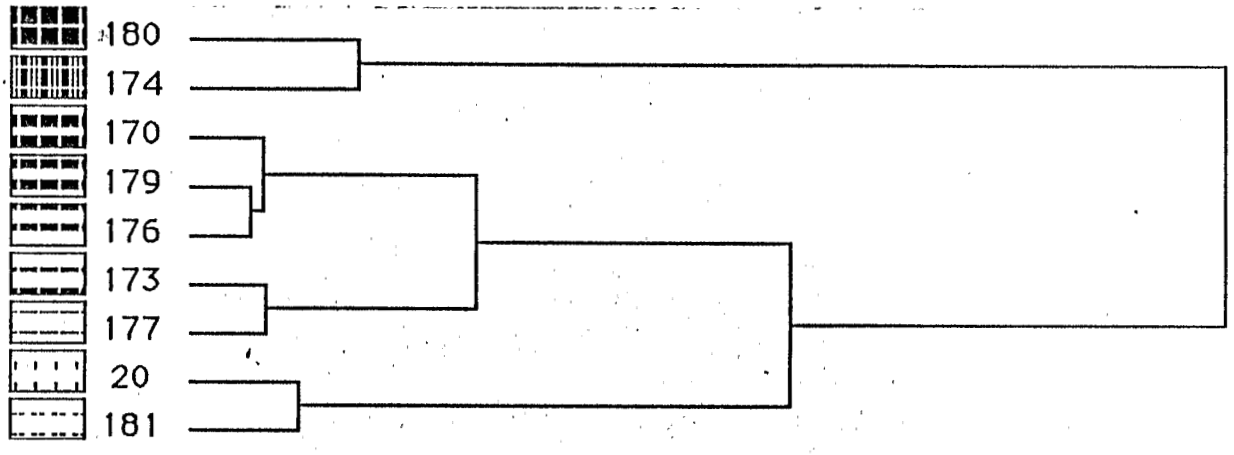


Si le choix des grisés n'apparaît pas satisfaisant pour traduire les contrastes des activités économiques entraînant une forte consommation de gazole en été ou en hiver et en automne avec les transitions que cela suppose, alors l'utilisateur peut remplacer les trames non adéquates par d'autres. Par exemple, il va remplacer la trame 179 de valeur trop faible par une trame de valeur plus élevée et ne confirmera donc pas le choix des trames proposées. Pour ce faire, il dispose d'une palette de trames (graphique ci-dessous) où sous chaque rectangle tramé sont écrits 3 nombres.
q = 4 c = 179 t = 9
ce qui signifie que cette classe est la neuvième de la partition choisie sur l'ensemble des départements, qu'elle porte dans la CAH le numéro 181 et qu'on lui avait attribué la trame n° 9.

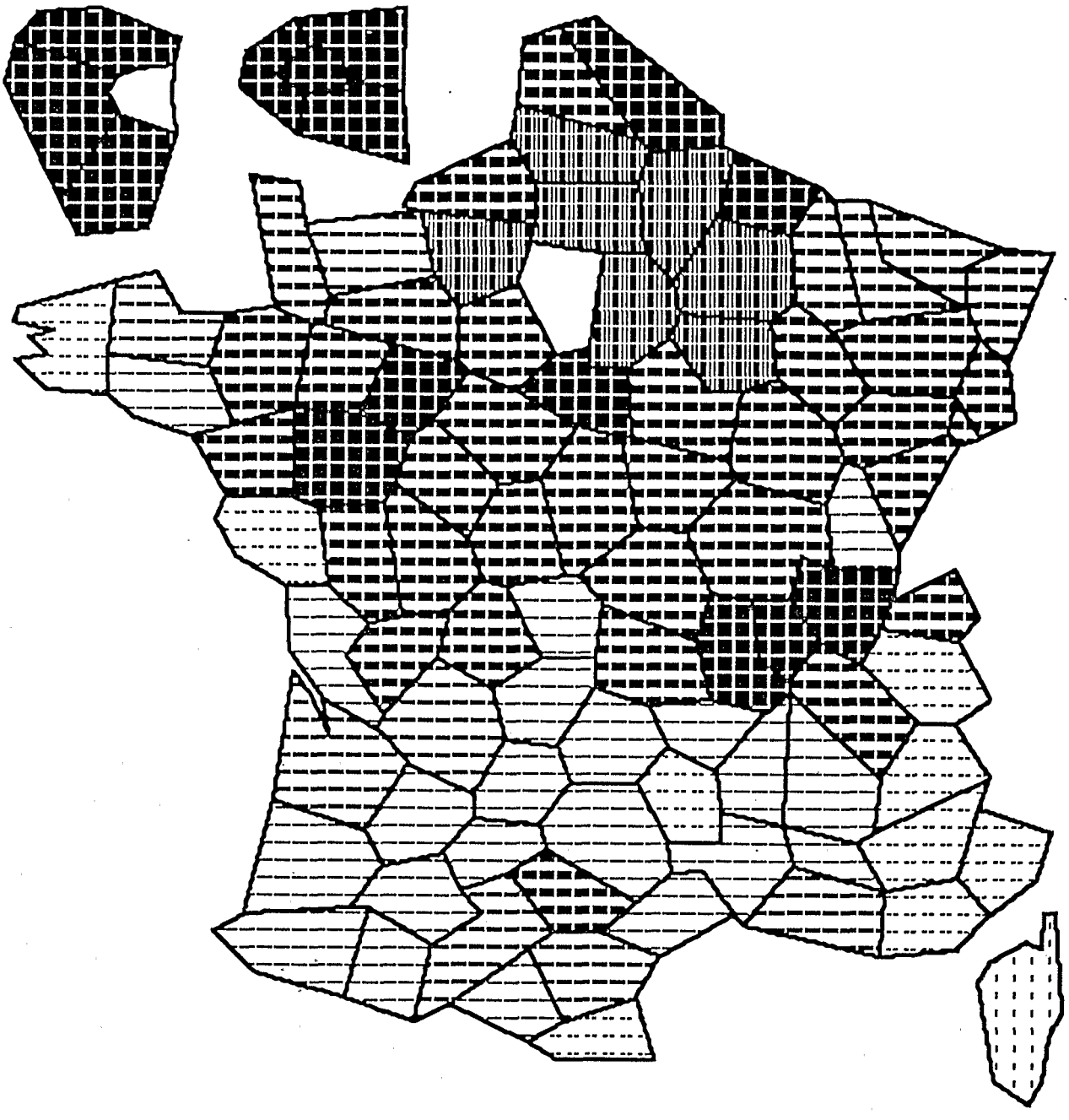


on modifiera la trame de la classe q= 4
la trame demandee pour la classe 4 est t=4
faut il modifier une autre trame oui(O) ou non(N)

Pour procéder aux modifications jugées souhaitables, l'utilisateur s'appuie pour le choix des grisés sur l'éventail des gammes proposées et demande la trame souhaitée.
Le dialogue de choix des trames se poursuit si d'autres substitutions semble judicieuse à l'utilisateur.
Les trames disposées dans la palette comportent 3 gammes de grisés, d'épaisseur de traits variant selon l'horizontale, la verticale et ces deux orientations - en outre les trames proposées au départ, avant tout dialogue sont attribuées systématiquement de 1 à 9. Ainsi, l'attribution d'office traduit un phénomène d'intensité que l'utilisateur à son choix pourra moduler selon le phénomène à représenter et la partition choisie.



La présentation de l'arbre des classes illustrée par des caissons tramés est indispensable car elle constitue, après réduction, la légende de la carte (cf graphique ci-dessus). L'information figurant sur cet arbre permet de confirmer la validité des choix effectués mais qui pourront de nouveau être modifiés après affichage de la carte par un passage dans de nouvelles étapes du dialogue que nous sommes en train de parcourir. La valeur et le figuré des grisés traduisent les associations de la hiérarchie.
La taille de l'écran du Macintosh +, à la différence du Macintosh 2, ne permet pas l'affichage à une échelle satisfaisante de la carte entière. La partie nord de la carte s'affichera alternativement avec la partie sud, affichage comportant une bande de recouvrement suffisante pour faciliter la lecture à l'écran ou la réalisation du montage des sorties graphiques nord et sud pour constituer une carte entière. L'épaisseur du trait (variant de 1 à 3) est à préciser, pour le cartographe trait fin : 1.
La partie nord de la carte affichée à l'écran comporte au nord-ouest, en carton, la grande et la petite couronne de Paris.



- Le programme "Carthag"

La différenciation des grisés par leur valeur et leur figuré, traduit sur la carte imprimée la différenciation saisonnière de la consommation du gazole. Nuancée dans la partie nord, l'opposition apparaît plus brutale dans la partie sud, encore que quelques régions industrialisées s'individualisent par des zones sombres quadrillées et que quelques départements se caractérisent par des lignés de valeur moyenne. Dans la partie sud, les zones claires prédominent illustrant la prédominance de la consommation estivale liée aux activités de vacances.
Au cas où l'utilisateur ne serait pas entièrement satisfait de la présentation cartographique obtenue, il peut entamer dès lors, un nouveau parcours, en reprenant le dialogue à partir des résultats de l'A.F.C.

Accès au programme Carthag sans passer par la chaîne d'Analyse des Données.

L'utilisation du programme Carthag peut se faire sans utiliser la chaîne de traitement de l'Analyse des Données. On peut aussi désirer utiliser l'ordinateur Macintosh pour représenter cartographiquement les résultats d'analyses de données effectuées sur un autre ordinateur.

Si l'on borne son ambition à la réalisation de la carte sans espérer avoir également l'affichage sur un plan des classes d'unités territoriales et classes de variables avec leurs coordonnées, comme nous l'avons précédemment obtenu, alors l'utilisateur crée un seul fichier intermédiaire décrivant la partition adoptée pour l'ensemble des départements.
Ce fichier sera créé en 2 étapes, d'abord création d'un fichier de type texte, assez semblable au fichier des sommets et à celui des unités territoriales (fichiers "stx" et "utx" créés pour le fond de carte), ensuite en faisant agir sur ce fichier de type text un programme de lecture qui le vérifiera (c'est-à-dire en vérifiera la cohérence interne) et créera le fichier numérique utilisé par le programme de cartographie.
Ouvrons donc un exemple de fichier cartographique.
Il a reçu pour nom "videicqktx". "Vide" joue dans cette cartographie le même rôle que "gazole" précédemment. Pourquoi ce nom ? pour nous rappeler que cette partition ne se base sur aucune donnée. Le suffixe "icqk" est celui attribué par l'ordinateur au fichier intermédiaire contenant la description de la partition quand elle est réalisée par le programme "CHrbz". Enfin le suffixe "tx" est ajouté parce qu'il s'agit d'un fichier de texte.
Ce fichier a pour titre "videickqtx une partition arbitraire des départements français". La mention entre accolades indique ce que l'on a fait : on a mis le numéro de la classe (-numéro d'ailleurs arbitraire par exemple 15 ; dans notre exemple) ajouté au nombre 20000 (20115), puis mis le cardinal de la classe (le nombre de départements contenu dans cette classe) ajouté au nombre 10000 (10025) et ensuite les numéros des départements individuels, ici pris successivement de 1 à 95. Comme toujours, il importe de savoir, ce qui est pertinent dans le tableau, de ce qui ne l'est pas.
Ce qui est essentiel, c'est la succession des nombres, d'abord le nombre 7 (7classes) ensuite les groupes de nombres pour chaque classe : un nombre commençant par 20000 indiquant le numéro donné à la classe, -numéro arbitraire mais qui ne doit pas dépasser 999, ensuite le nombre des individus de la classe ne pouvant dépasser le nombre total d'unités territoriales considérées, -nombre ajouté à 10000, et enfin la liste des numéros de ces unités.
Les allers à la ligne ne jouent aucun rôle. Quant aux indications figurant après le chiffre 7 (classe et une phrase) pourvu qu'elles ne comportent aucun chiffre, elles ne gênent ni ne servent au déroulement du programme. Seuls comptent les nombres et les blancs auxquels s'ajoutent éventuellement des lettres qui séparent ces nombres.
Ce format avec des 20000 et des 10000 permet au programme de lecture qui va transformer ce fichier texte en un fichier numérique, d'effectuer un certain nombre de vérifications et de donner éventuellement à l'utilisateur des indications sur les incohérences que peut recéler son fichier.

Conclusion

Dans notre étude de la régionalisation agricole de la Sierra équatorienne, nous avons appliqué les méthodes de l'A.F.C. et de la C.A.H. à un ensemble de 238 circonscriptions administratives (paroisses). Les représentations cartographiques résultant de cette application, de grand format, à la numérisation des contours précise, dont le choix des teintes a été effectué d'après l'arbre de longueur minima, ont été réalisées en collaboration avec l'Institut Géographique National (I.G.N). De telles possibilités ne sont évidemment pas à la portée d'un micro-ordinateur Toutefois la représentation d'information statistique ventilée par circonscription administrative, peut se faire très rapidement en noir et blanc avec une bonne qualité de tracé, si on utilise une imprimante usuelle, excellente si l'on a accès à une imprimante laser.
Qu'il s'agisse de chercheurs, spécialistes, décideurs, l'utilité de tels documents est manifeste. Une représentation cartographique équivalente faisant la synthèse d'un ensemble de variables demande un temps certain de réflexion, si l'on n'a pas recours à l'Analyse des Données et la réalisation de la carte requiert, elle aussi, beaucoup de temps. Or l'automatisation permet en quelques heures de sortir et de comparer de nombreux essais avant d'arrêter un choix définitif. Il nous a paru utile, à travers un exemple, de présenter la conjugaison de ces formes nouvelles de l'expression documentaire avec la tradition cartographique.

Bibliographie.

BENZECRI (J.-P. & F.). Pratique de l'Analyse des Données, Analyse des correspondances. Exposé élémentaire, Dunod, Paris, 424 pages.

BENZECRI (J.-P. & F.) 1986. Pratique de l'Analyse des Données en Economie. Dunod, Paris, 534 pages.

Les Cahiers de l'Analyse des Données. Dunod, Paris.