CONSTRUIRE SES CARTES : LE DÉVELOPPEMENT D'OUTILS STATISTIQUES INTERACTIFS INTÉGRÉS À UN SYSTÈME D'INFORMATION GÉOGRAPHIQUE

Benoît OGIER
UPRESA 6063
Laboratoire Modélisation et Traitement Graphique
Université de Rouen

Résumé

Les systèmes d'exploitation des micro-ordinateurs proposent aujourd'hui des interfaces conviviales. En outre, certains micro systèmes d'information géographique offrent des possibilités intéressantes de développement par l'intermédiaire de « plug ins », dans les langages de programmation les plus courants (Pascal, C, Delphi...). C'est sur ces bases que nous avons pu compléter les fonctionnalités d'un micro système d'information géographique (requêtes spatiales, calcul d'itinéraire, gestion de données, cartographie...), par une série de modules statistiques et graphiques.

Ces outils mettent à la disposition de l'utilisateur toutes les informations statistiques et représentations graphiques nécessaires à une analyse complète des données. En effet, les modules sont liés entre eux, ce qui permet d'aborder une étude sous plusieurs angles simultanément et de comparer les résultats de différentes méthodes. De plus, une mise à jour automatique de la carte s'effectue dès que l'on procède à une mise en classes à l'aide d'un des modules. On est donc en présence d'une complète interactivité système d'information géographique-cartes-statistiques.

Mais on peut aller plus loin et envisager l'expérimentation d'autres outils graphiques, notamment pour l'exploitation des résultats d'analyses factorielles.

Abstract

Nowadays, microcomputers operating systems offer user-friendly interfaces. Morever, some micro-GIS have interesting development possibilities through plug ins, in the common programming languages : Pascal, C, Delphi, etc... On this basis, we have complemented functions of one micro-GIS (spatial requests, itinerary calculation, data managing, cartography) with a set of statistical and graphical modules.

These tools give to users all statistical informations and graphical representations that are necessary for a comprehensive data analysis. Interactive links between modules allow a global approach, and the comparison and merging of different methods of analysis. Lastly, the associated map is triggered anytime clusters of individuals are defined. Indeed, we do have a complete interactivity between GIS, maps, and statistical modules.

However, we plan to go further in experimenting of other graphic tools, especialy in the exploitation of factor analysis results.

Mots-Clés

Analyse factorielle, interactivité, matrice de Bertin, outil d'analyse spatiale, système d'information géographique

Key-Words

Factor analysis, interactivity, Bertin matrix, spatial analysis tool, geographical information system

Depuis une douzaine d'années, le développement de la micro-informatique et l'amélioration des performances des micro-ordinateurs, ont mis à la disposition de la communauté géographique des outils qui, grâce à leurs capacités croissantes et à la simplification des interfaces, permettent d'accroître la quantité de données traitées, d'automatiser et de simplifier un certain nombre de tâches. Le développement de logiciels de traitement des données, et celui des systèmes d'information géographique (SIG), sur station, et sur micro-ordinateur, ont facilité la diffusion de l'informatique dans le monde de la géographie universitaire. Le

développement spécifique des systèmes d'information géographique a contribué à populariser les approches géographiques, chez les utilisateurs non universitaires (entreprises, administrations, collectivités locales...), et les développeurs de systèmes d'information.

Malgré cette diffusion de l'outil informatique, et l'intégration de plus en plus courante, dans les travaux de recherche, de données spatialisées, on constate que les éditeurs de logiciels ne proposent pas toujours de solutions combinant système d'information géographique et analyse des données. De son côté, le monde universitaire est riche de réflexions et de travaux sur les potentialités offertes par les systèmes d'information géographique, mais les développements réellement proposés sont assez rares. L'intérêt d'un outil de ce type est donc bien réel, puisqu'il permet de mieux lier les données géographiques (position dans l'espace, forme...) aux données statistiques (population, cultures portées par une parcelle...). Ainsi, on pourrait pratiquer une véritable analyse spatiale, au lieu de simplement superposer les résultats d'une analyse des données dans un système d'information géographique, ou d'utiliser un système d'information géographique pour produire des données que l'on analysera ensuite. Un travail important doit dès lors être effectué, afin d'apporter des solutions opérationnelles, impliquant toutefois d'accepter un investissement conséquent en programmation.

Après avoir développé les principes de base d'un tel système, nous exposerons les développements informatiques envisagés, pour terminer sur quelques exemples d'utilisation.

1. Principes de base

1.1. Mieux intégrer le traitement des données aux systèmes d'information géographique

Il n'existe guère de logiciels intégrant à la fois des possibilités d'analyse des données et des éléments de système d'information géographique. Pourtant, la possibilité de traiter directement les données à l'intérieur d'un système d'information géographique simplifierait énormément le travail des utilisateurs. Les procédures d'importation et d'exportation disponibles dans la plupart des systèmes d'information géographique et rendues nécessaires par le traitement séparé des entités géographiques et de leurs données, sont en général lourdes à utiliser, représentent un travail important dès que les bases sont tant soit peu complexes, et multiplient les risques d'erreur.

L'intérêt d'un tel outil est cependant beaucoup plus étendu et dépasse même la somme des possibilités offertes par des applications séparées. Il faut en effet l'appréhender comme une plate-forme permettant de réaliser une véritable analyse spatiale, c'est-à-dire, une analyse prenant en compte l'aspect statistique d'un problème tout en intégrant les préoccupations géographiques. Il s'agit donc bien de mettre au point un outil permettant d'appréhender les entités géographiques dans leur globalité pour une analyse plus complète.

Les développements à réaliser concernent donc l'analyse statistique des données. Les traitements multivariés, les analyses factorielles ou les classifications hiérarchiques, du fait des possibilités de synthèse qu'ils apportent, constituent la première étape des développements à réaliser. Ces algorithmes relevant du domaine public, il est aujourd'hui très simple de trouver une bibliographie complète sur les méthodes à mettre en oeuvre. Des outils habituels d'analyse uni-variée et bi-variée sont, bien entendu, indispensables pour l'étude simple et la critique préalable des données.

1.2. Spécificités de l'interface utilisateur

La plupart des micro-ordinateurs proposent, aujourd'hui, des interfaces graphiques intégrant les notions d'interactivité et de réactivité. La majorité des systèmes d'information géographique développés sur ces plates-formes intègrent donc, eux aussi, ces notions, et permettent à l'opérateur de sélectionner des objets

par simple clic à la souris de manière immédiate et quasi instinctive. Ces possibilités offertes par les interfaces graphiques sont à la base de nos développements, et permettent une simplification du travail d'analyse. On constate, en effet, que, malgré la puissance de méthodes telles que les analyses factorielles des correspondances (AFC) ou les analyses en composantes principales (ACP), l'exploitation des nuages factoriels (ou mapping) issus de ces analyses, reste un point assez délicat quand on recherche des types d'individus cohérents. L'interface utilisateur doit donc s'attacher à simplifier l'exploitation de ces nuages en permettant, par exemple, de visualiser simultanément différents plans factoriels, et d'y localiser un individu ou un groupe d'individus par un mécanisme de sélection. L'aide à l'analyse peut même aller plus loin, en offrant la possibilité de réaliser simultanément plusieurs traitements, puis de rechercher des noyaux stables d'individus. Cette approche est préconisée, en général, pour dégager des groupes d'individus à travers une classification hiérarchique et une analyse factorielle, par exemple.

L'interface utilisateur doit, de plus, intégrer un mécanisme liant les modules statistiques à la carte, de manière à offrir une aide à la cartographie en premier lieu, mais aussi pour mettre à la disposition de l'utilisateur des éléments de compréhension des mécanismes régissant les liens entre individus statistiques (ou objets géographiques).

Ces modules doivent offrir un ensemble de fonctions permettant de réintégrer les résultats d'une analyse (coordonnées des individus sur différents axes, numéros de groupes attribués par l'utilisateur...) dans le système d'information géographique de manière à pouvoir réutiliser ces données, dans une analyse ultérieure. Il est en effet courant d'effectuer une classification ascendante hiérarchique ou encore une partition par centres mobiles sur les coordonnées des individus d'une analyse factorielle des correspondances ou d'une analyse en composantes principales. On peut, de plus, utiliser les coordonnées des individus et des variables pour effectuer un pré-traitement d'une matrice graphique.

Enfin, il faut mettre en œuvre des solutions simples pour l'exportation des données et des nuages factoriels vers d'autres logiciels (tableurs, système de gestion des bases de données, traitements statistiques et dessin assisté par ordinateur), notamment par "copier-coller" et par génération de fichiers textes ou d'images.

2. Développements informatiques

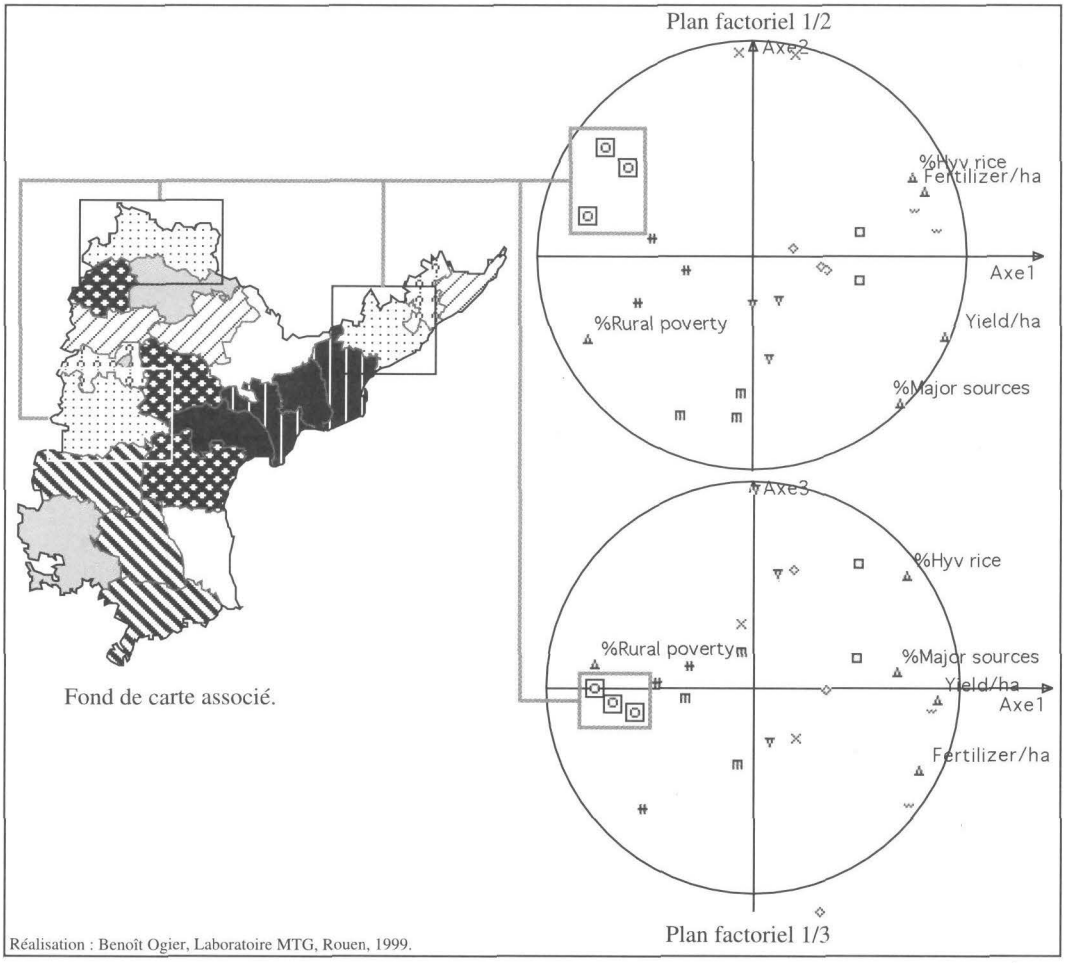
2.1. Choix de la plate-forme de développement

Le département de géographie de l'université de Rouen dispose de plusieurs logiciels de système d'information géographique: ArcInfo, ArcView, MapInfo, MacMap® et GeoConcept. MacMap®, système d'information géographique vectoriel possédant déjà quelques outils de traitement des données, s'est avéré être la meilleure solution pour ce travail. En effet, il offre des possibilités de représentations cartographiques étendues et une interface utilisateur conviviale. En outre, il est possible de programmer des extensions à l'aide d'un "kit de développement", commercialisé en langages Pascal et C. De plus, il est utilisé au sein des unités de recherche (MTG[1], LEDRA[2]), mais aussi en libre accès pour tous les étudiants en géographie. Enfin, une matrice graphique (module "Bertin"), créée en collaboration avec la société Décision Graphics[3], des chercheurs de l'université de Rouen et la société Klik Développement[4], permet déjà d'effectuer des typologies à partir des données gérées par MacMap®. Nous nous proposons d'aller plus loin en développant des outils permettant une véritable analyse des données.

Avant tout, il était nécessaire de maîtriser le langage Pascal, la programmation sur Macintosh et les fonctions du "kit de développement". D'ores et déjà, les prototypes des différents modules fonctionnent et permettent de réaliser des analyses uni-variées, bi-variées, des analyses en composantes principales, des analyses factorielles des correspondances et des classifications ascendantes hiérarchiques. Une nouvelle version de ces développements, stabilisée et transcrite en langage C, sera bientôt fonctionnelle.

2.2. Fonctionnement général

Les modules développés ont été construits sur un concept orienté objet, et leurs structures sont standardisées, ce qui permet une assez large réutilisation du code. Cela se traduit pour l'utilisateur par une interface standardisée et des fonctionnements similaires. Un premier tableur permet de sélectionner et de visualiser les données que l'on désire intégrer à l'analyse. Une fois le choix des données effectué, le calcul peut être lancé. A l'issue de ce calcul, différents indicateurs sont proposés par l'intermédiaire de tableurs spécifiques, et des représentations graphiques permettent d'effectuer des mises en classe des individus, mettant à jour le fond de carte associé, si nécessaire. Les individus peuvent êtres pointés à la souris, soit sur les représentations graphiques, soit dans les tableurs, ce qui provoque leur sélection, donc leur localisation, sur la carte et dans d'autres analyses effectuées simultanément. Inversement, la sélection d'un objet sur la carte entraîne une mise à jour automatique de la sélection dans les tableurs et représentations graphiques. Le processus de sélection des objets (informatiques et géographiques) proposé par MacMap® est donc généralisé, et l'utilisateur peut ainsi employer les modules en interaction avec la carte et avec les autres modules éventuellement utilisés (fig. 1). De plus, les représentations graphiques sont complétées par des outils



La définition de groupes est facilitée par les possibilités de repérage des individus, simultanément, sur différents plans factoriels, et sur la carte associée. Ce repérage peut s'effectuer par sélection des objets, et par attribution d'icônes spécifiques à chaque groupe. Une mise à jour automatique de la carte est réalisée à chaque création de groupe.

Figure 1 - Détection et localisation de groupes d'individus

permettant de grossir ou de réduire l'affichage, et de se déplacer. On peut ainsi explorer les nuages factoriels trop compacts. Il est aussi possible d'effectuer plusieurs analyses simultanément, par exemple, une analyse en composantes principales et une classification ascendante hièrarchique, ce qui permet par comparaison de vérifier la pertinence des discrétisations. On peut, enfin, obtenir différents indices (moyenne, écart-type...) sur les variables étudiées à l'aide des modules uni-variés et bi-variés. Précisons, pour terminer, que toutes les données produites par ces modules (y compris les représentations graphiques) sont récupérables par exportation ou copier-coller, et que celles relatives aux individus (coordonnées, contributions...) peuvent être réintégrées directement à la base de données. Elles peuvent, de plus, être enregistrées et réexploitées par la suite.

3. Exemples d'utilisation pour l'aide à l'analyse

3.1. Exemples de croisements de méthodes

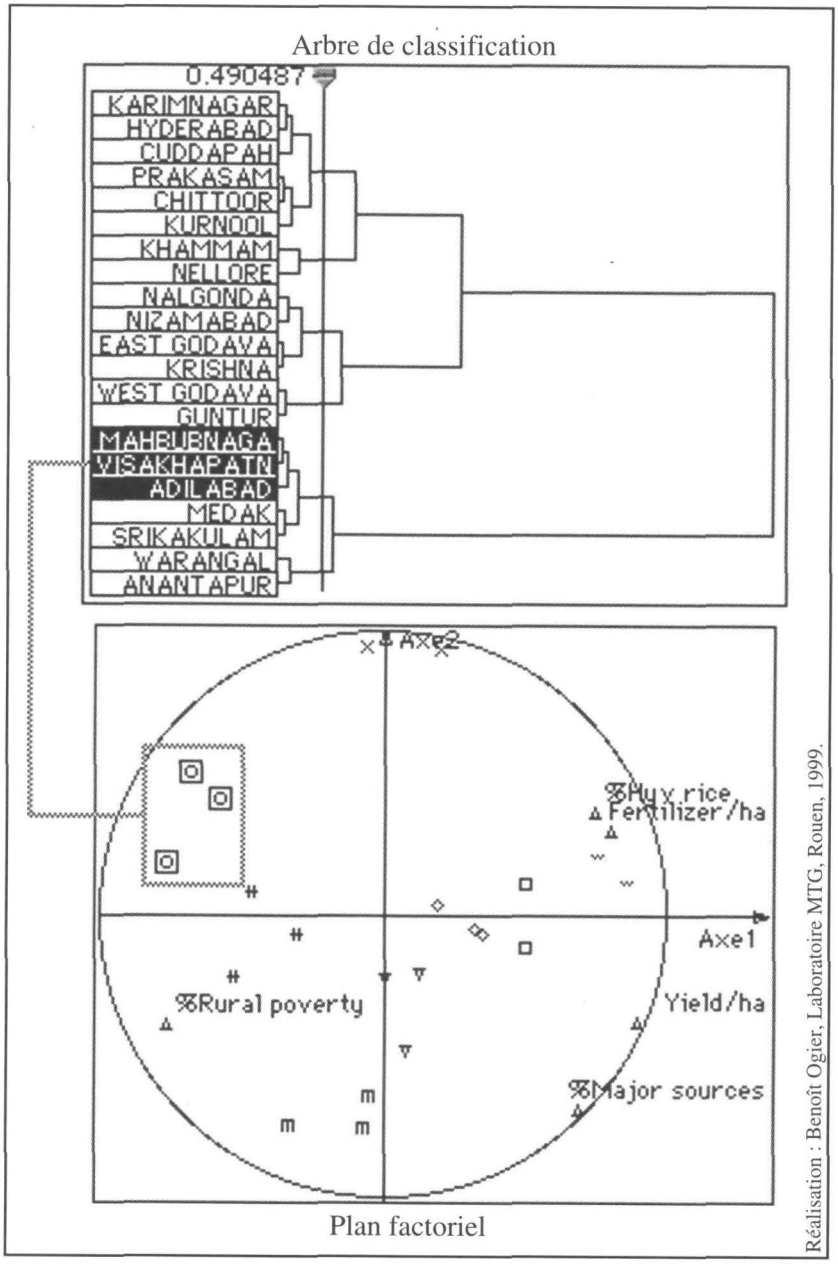
3.1.1. Aide à l'exploitation des résultats d'analyse factorielle par recoupement de plans

Lors de la constitution de typologies, la comparaison des positions respectives des individus, ne serait-ce que sur deux plans, devient très vite un exercice fastidieux quand le nombre d'individus est élevé. Il nous est donc apparu nécessaire de faciliter au maximum ces comparaisons, en permettant à l'utilisateur de repérer simplement les individus sur les différents plans. Leur différenciation se fait de deux manières. Tout d'abord, avant constitution des groupes, il est possible de les localiser par simple sélection. Un individu sélectionné sur un plan factoriel l'est aussi dans tous les autres plans définis. D'autre part, l'attribution d'un numéro de groupe à un ou plusieurs individus entraîne une modification de leur représentation dans les différents plans. Une série de 256 icônes permet, en effet, de définir autant de groupes. L'utilisateur peut ainsi visualiser immédiatement la cohérence d'une classification (fig. 1).

La liaison entre la carte et les différents modules de traitement apporte aussi une aide pour la définition des groupes. Ce lien est réalisé, d'une part, grâce aux mécanismes de sélection, d'autre part en fonction des choix de l'utilisateur, par une mise à jour automatique du fond de carte à chaque attribution d'individus à un groupe. L'interactivité permet ainsi de véritablement construire ses cartes à partir des résultats d'une analyse et en fonction des connaissances de l'utilisateur.

3.1.2. Exploitation conjointe d'analyse factorielle et de classification ascendante hiérarchique

Il est conseillé, en général, de recouper les méthodes statistiques afin de détecter les noyaux stables d'individus. L'utilisation la plus fréquente de cette méthode est l'exploitation simultanée d'une analyse factorielle et d'une classification ascendante hiérarchique. Deux procédures peuvent alors être envisagées. La plus simple consiste à réaliser simultanément, sur le même groupe d'individus et de variables, une analyse factorielle et une classification hiérarchique, puis de dégager des groupes d'individus à l'aide des deux méthodes. Ce travail est extrêmement simplifié, dans nos développements, par l'interactivité des modules et les processus de mise à jour des éléments sélectionnés (fig. 2). Une autre solution consiste à effectuer une classification ascendante hiérarchique à partir des coordonnées des individus sur les axes [5], soit pour utiliser des données hétérogènes, "normalisées" par des analyses factorielles, soit, en se limitant aux facteurs les plus représentatifs, pour éliminer le bruit résiduel expliquant les axes les moins structurants, ou pour accélérer l'agrégation des individus. Ici encore, ce travail est facilité par les possibilités de réintégration des données produites par les analyses factorielles.

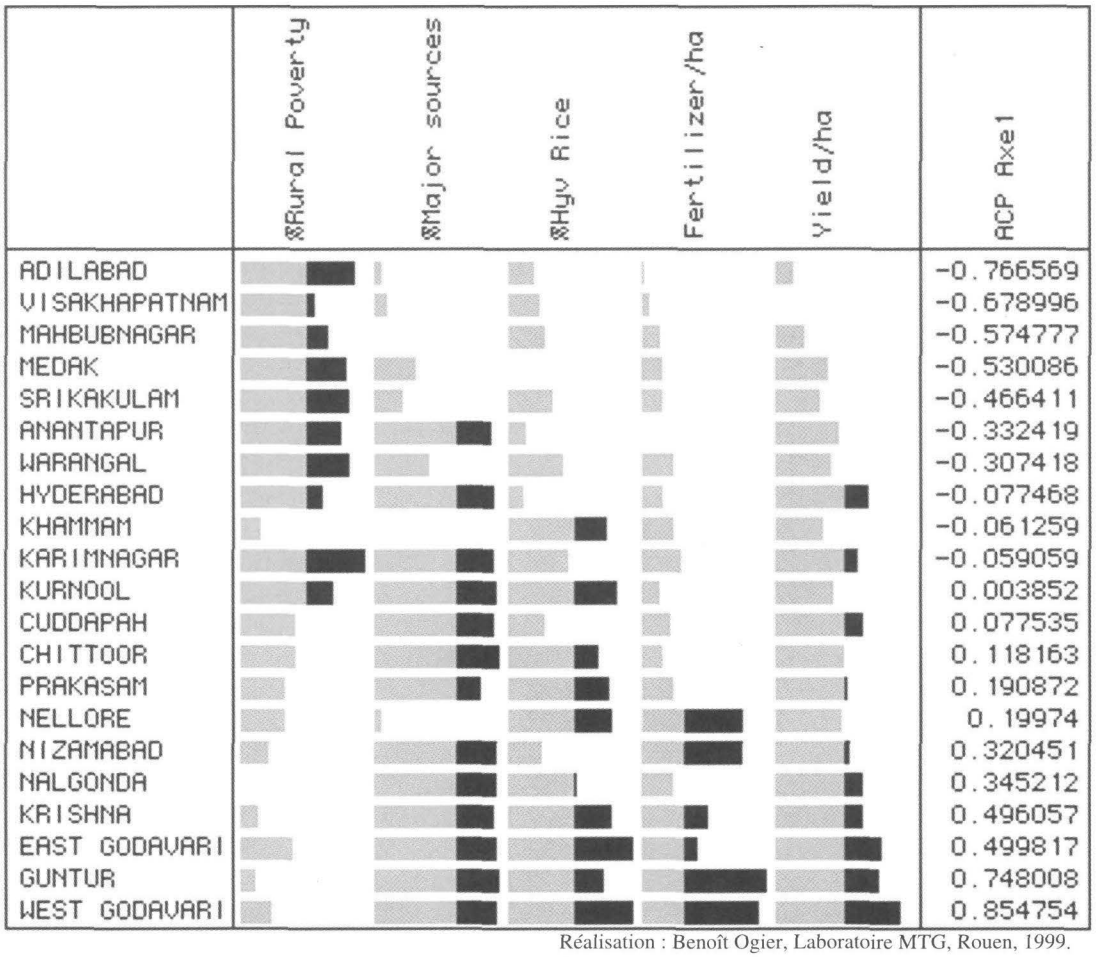


La détection des groupes peut être améliorée par l'utilisation conjonte d'une ACP et d'une CAH sur les mêmes données. Les mécanismes de sélection et de rafraîchissement de la carte permettent une meilleure visualisation de proximité entre les individus. La mise en œuvre d'une CAH à partir des coordonnées des individus sur les axes factoriels d'une CAP ou d'une AFC est, de même, tout à fait réalisable.

Figure 2 - Détection des groupes par recoupement de méthodes : analyse en composantes principales – classification ascendante hiérarchique

3.1.3. Aide à l'organisation des matrices graphiques

Il existe, sur MacMap®, un module permettant d'effectuer des typologies à partir de matrices graphiques (type Bertin). La mise en forme manuelle de ces matrices devient, en général, assez fastidieuse dès que l'on atteint la centaine d'individus et quelques dizaines de variables. Les limitations imposées par les écrans viennent, de plus, compliquer un peu plus le travail. Le développement d'outils d'analyse factorielle peut permettre un pré-traitement de ces matrices de manière extrêmement simple, rapide, et en général, assez efficace. Il suffit, en effet, dans un premier temps, de réaliser une analyse factorielle sur les données et les individus que l'on désire intégrer à une matrice graphique. Les coordonnées des individus sur l'axe principal sont ensuite intégrées dans un champ de la base de données. Ce travail effectué, on constitue la matrice en y intégrant, temporairement, le champ dans lequel on a stocké les coordonnées des individus et l'on effectue un tri sur ce champ pour obtenir un premier stade d'organisation visuelle. On organise ensuite les variables entre elles en fonction de leurs coordonnées sur l'axe principal. Le stade de pré-traitement est alors atteint, et il ne reste plus qu'à effectuer les mise en formes ponctuelles (fig. 3). Ce type de pré-traitement est proposé dans le



Réalisation : Benoît Ogier, Laboratoire MTG, Rouen, 1999.

Le pré-traitement est obtenu par tri des coordonnées des individus et des variables sur l'axe principal d'une analyse factorielle. Le champ contenant les coordonnées des individus n'est utilisé que pour effectuer le tri. L'organisation finale doit être achevée manuellement.

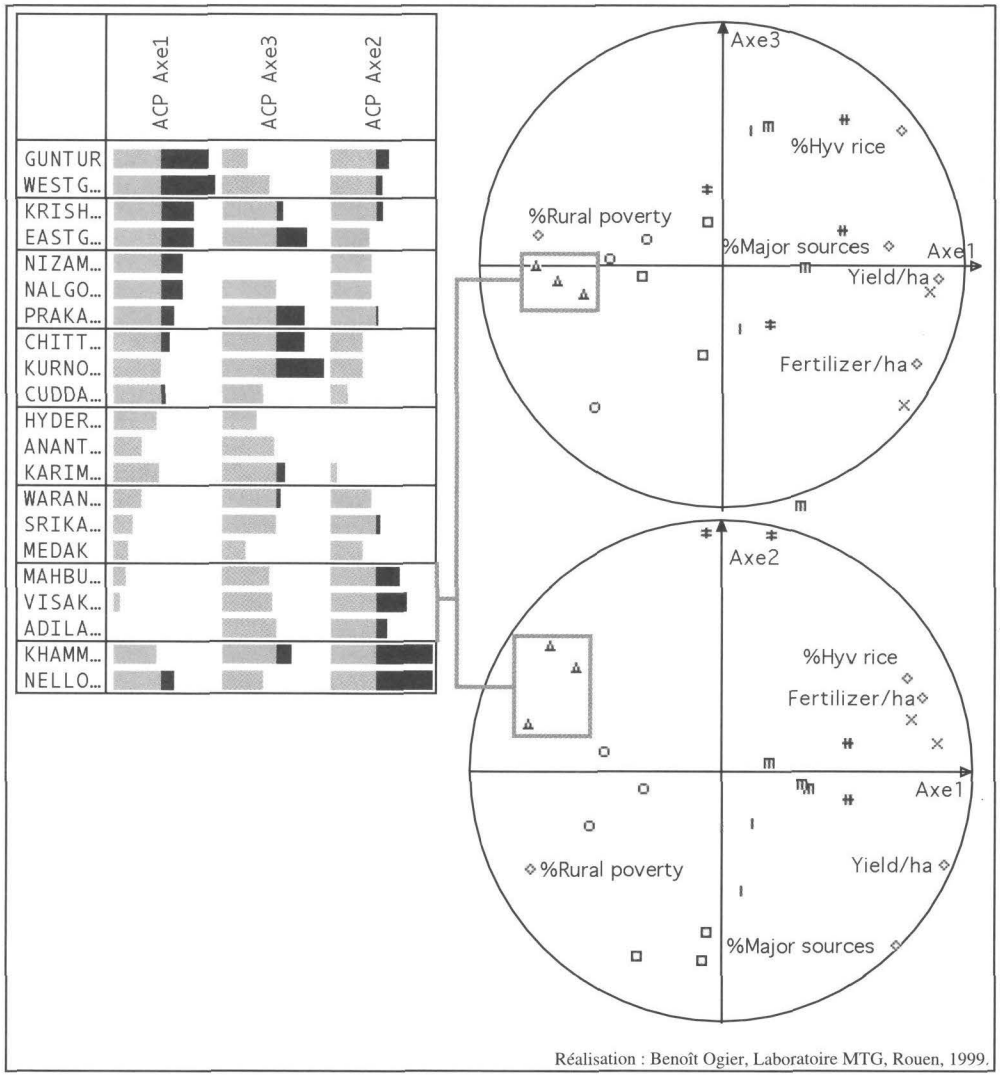
Figure 3 - Pré-traitement d'une matrice graphique par analyse en composantes principales

logiciel AMADO développé par Alban Risson (CISIA[5]). Le fonctionnement du module classification ascendante hiérarchique ne permet pas, pour l'instant, d'obtenir des données pour l'organisation des matrices graphiques, mais cette possibilité doit être explorée. Notons toutefois que l'on ne dispose pas, dans ce cas, de données pour l'organisation des variables entre elles.

3.2. Proposer d'autres outils pour l'aide à l'analyse

3.2.1. Aide à l'exploitation des résultats d'analyse factorielle par matrices graphiques

La matrice Bertin pourrait apporter une aide à la constitution de typologies sur la base des résultats d'analyses factorielles. La détection de groupes d'individus est généralement problématique dans les cas où



La réalisation d'une matrice graphique sur les coordonnées des individus d'une analyse factorielle permet de représenter simultanément plus de deux axes significatifs pour la recherche des groupes.

Figure 4 - Détection des groupes par recoupement de méthodes : analyse en composantes principaleset matrice graphique

le nombre d'axes à prendre en compte est supérieur à deux. Les recoupements de plans peuvent alors devenir multiples, et d'autant plus longs à réaliser que les individus sont nombreux. Les graphiques en trois dimensions sont difficiles à utiliser, aussi bien sur écran que sur papier. Une matrice graphique peut alors devenir le moyen de représenter virtuellement un espace factoriel comprenant autant de dimensions que d'axes signifiants. En intégrant à cette matrice les coordonnées des individus sur ces axes, on obtient une représentation plane et compréhensible de l'espace factoriel défini, à partir duquel on peut réaliser une mise en classes (fig. 4).

3.2.2. Aide à l'exploitation des résultats d'analyse factorielle par matrices trichromatiques

Nous explorons aussi les possibilités de représentation d'un espace factoriel en trois dimensions simulé par matrice trichromatique, fonctionnant de manière similaire aux matrices graphiques (AMADO, Bertin...). Les axes factoriels spécifiés par l'utilisateur sont codés sur les trois couleurs fondamentales en synthèse additive, le rouge, le vert et le bleu. Ces trois composantes, dites RGB, sont en effet utilisées par les écrans cathodiques pour afficher toutes les nuances voulues. Le principe de codage est simple, puisque pour chacun des axes retenus, on calcule la distance entre chaque individu et chaque variable, puis on procède à un étalement des valeurs ainsi obtenues sur l'espace des valeurs RGB (codées sur un entier non signé sur le Macintosh : 0 - 65535). Les axes sont ensuite affectés à une couleur, et l'on obtient, par synthèse, une image à organiser manuellement, et à l'aide de tris sur les différents axes. Cet outil doit être appréhendé comme une aide supplémentaire conjointe à l'utilisation des représentations graphiques, un moyen de contrôle en quelque sorte, sa mise en forme manuelle s'avérant laborieuse. Un point doit être précisé : dans le cas d'une analyse en composantes principales, la représentation des individus et des variables sur un graphique est virtuelle puisque les valeurs ne sont pas du même ordre : on représente des vecteurs (les variables) et des points (les individus). Dès lors, sur un graphique, la distance entre un individu et une variable n'a pas de sens pour elle-même, et l'explication d'un groupe d'individus se fait par l'intermédiaire de l'axe, lui même expliqué par les contributions d'une ou plusieurs variables. La distance calculée sur laquelle est fondée cette matrice n'a donc pas de sens. Cependant, la représentation trichromatique gomme la signification que l'on pourrait accorder, à tort, à cette valeur, car on ne dispose pas d'une relation quantifiée entre individus et variables. On ne peut que comparer et assembler des profils pour obtenir une image synthétique. On ne tire donc pas de conclusion directe de la distance entre individus et variables, pas plus qu'on ne le fait dans l'exploitation d'un nuage factoriel classique.

Les modules présentés ici ne sont que des prototypes développés depuis juillet 1998 et de nombreuses améliorations doivent encore y être apportées. Des fonctions complémentaires doivent de plus être développées. Dans le domaine de l'interface utilisateur, d'autres travaux sont aussi à envisager, pour apporter une aide à l'analyse encore plus complète. Enfin, pour l'avenir, et en fonction de l'impact de ces travaux, il pourrait être envisageable de développer de nouveaux modules statistiques ou orientés vers le Data Mining.

L'intérêt de la mise au point de tels outils ne relève donc pas du gadget ou de la simple commodité. Il faut bien au contraire les considérer comme de nouvelles possibilités au service de l'étude de données spatialisées. Ces outils devraient donc permettre d'améliorer l'analyse en multipliant les recoupements d'informations et de méthodes, grâce aux liaisons entre les différents modules.

Références bibliographiques

[1] BEGUIN M., PUMAIN D., 1994 : *La représentation des données géographiques : statistiques et cartographies*, Paris, Armand Colin, 192 pages

[2] BOUROCHE J.-M., SAPORTA G., 1980 : *L'analyse des données*, Paris, PUF, 126 pages

[3] CIARLET P.G., 1998 : *Introduction à l'analyse numérique matricielle et à l'optimisation*, Paris, Dunod, 279 pages

[4] DROESBEKE J.-J., 1988 : *Eléments de statistiques*, Bruxelles, Editions de l'Université de Bruxelles, 446 pages

[5] ESCOFFIER B., PAGES J., 1998 : *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation*, 3e édition, Paris, Dunod, 284 pages

[6] JAMBU M., 1989 : *Exploration informatique et statistique des données*, Paris, Bordas, 505 pages

[7] LASCAUX P., THEODOR R, 1993 : *Analyse numérique matricielle appliquée à l'art de l'ingénieur*, Paris, Masson, 2 tomes, 636 pages

[8] VELU J., 1994 : *Méthodes mathématiques pour l'informatique*, Paris, Dunod, 452 pages

Notes

1 - Modélisation et Traitement Graphique

2 - Laboratoire d'Etude du Développement des Régions Arides

3 - SSII et Bureau d'études géographiques

4 - Klik Développement était la société créatrice du logiciel. Son éditeur actuel est la société Carte Blanche

5 - Centre International de Statistique et d'Informatique Appliquée