SIG ET ANALYSE EXPLORATOIRE

VERS DE NOUVELLES PRATIQUES EN GÉOGRAPHIE

Jean-Marc ORHAN
Equipe P.A.R.I.S., URA 1243 du CNRS
Paris

Résumé

L'offre actuelle dans le domaine des logiciels de type Système d'Information Géographique (SIG) est très vaste : plus d'une centaine existent et les utilisateurs travaillent couramment avec des bases de 300 000 objets. Les besoins dans les années à venir se situeront dans le domaine de la visualisation et de l'analyse de données.

L'intégration dans le SIG MacMap d'un module d'analyse statistique, et notamment la création d'un interface utilisateur, oblige à rechercher de nouveaux types de démarches d'analyses intégrant les possibilités offertes par les SIG. L'objectif est d'utiliser les méthodes de la cartographie pour représenter les résultats et faciliter leur prise en compte. On espère ainsi amener l'utilisateur à employer des outils statistiques, mais surtout à utiliser les capacités de gestion de l'espace du SIG pour créer de nouveaux espaces, nés des résultats de l'analyse.

L'objectif est donc de mettre au point des outils créant de nouveaux systèmes de coordonnées. D'un histogramme de fréquence à un plan factoriel ou à un nuage de points, on est à chaque fois en présence d'un de ces espaces recréés. Leur visualisation cartographique et leur mise en interaction devraient offrir aux géographes de puissants outils d'analyse.

Mots Clés

Analyse de données - Cartographie - Interactivité - Système d'Information Géographique - Visualisation

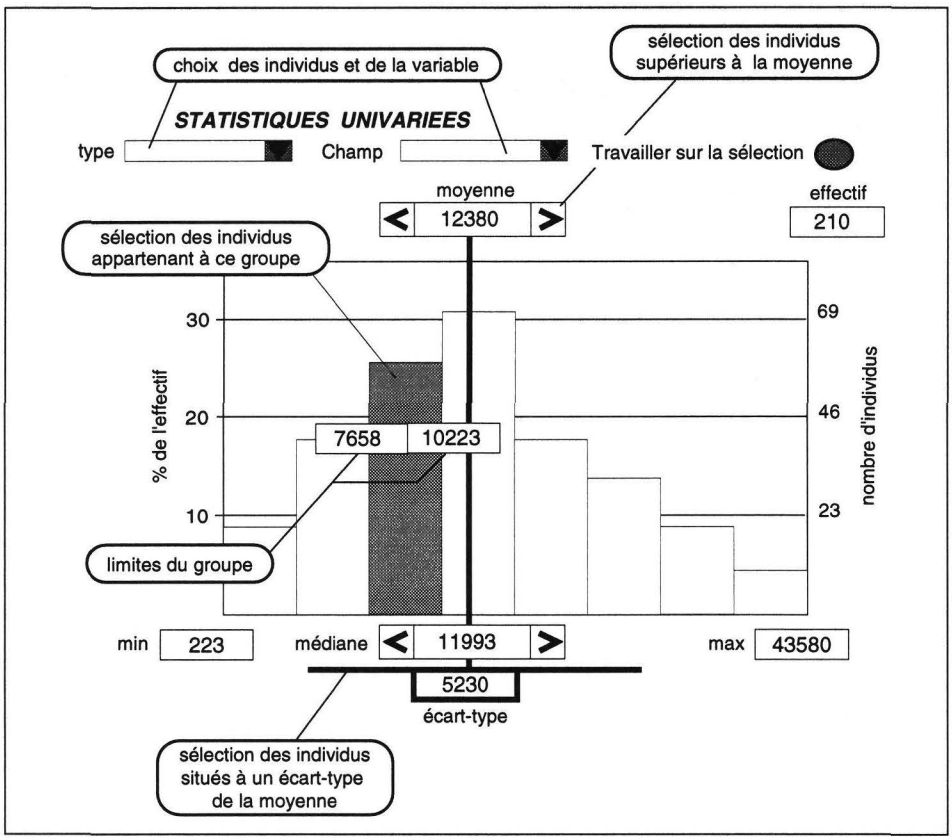
L'offre actuelle dans le domaine des logiciels de type Systèmes d'Information Géographique (SIG) est très vaste, plus d'une centaine de logiciels existent. Les SIG ont atteint leur maturité en matière de gestion de gros volumes de données, les utilisateurs travaillent couramment avec des bases de 300 000 objets. Cette capacité atteinte, les besoins qui surgiront dans les années à venir se situeront dans le domaine de la visualisation et de l'analyse de données. Les outils de l'analyse statistique sont actuellement extrêmement réduits dans les SIG ; l'utilisateur y pallie en intégrant des chaînes de traitement alliant SIG et logiciel de statistique (arcinfo-sas). Ce type de pratique empêche une prise en compte efficace de la composante spatiale des individus analysés, et ne permet aucune interaction entre le module d'analyse et l'espace géographique.

Nous avons le projet d'intégrer dans le SIG MacMap un module d'analyse de données. Son intégration, et notamment la création de l'interface utilisateur, nous oblige à rechercher de nouveaux types de démarches d'analyse intégrant les possibilités offertes par les SIG. La caractéristique spécifique de MacMap se situe dans le domaine de la visualisation ; il s'agit donc de montrer, dans un premier temps, que les règles de la graphique s'appliquent aussi bien à la cartographie qu'à l'analyse de données. L'objectif est d'utiliser les méthodes de la cartographie pour représenter les résultats des analyses et faciliter leur prise en compte. On espère ainsi amener l'utilisateur à employer des outils statistiques, mais surtout à utiliser les capacités de gestion de l'espace du SIG, pour créer de nouveaux espaces, nés des résultats de l'analyse. Ainsi, un graphique de régression ou un plan factoriel peut être considéré comme un espace et ses éléments être traités de manière cartographique.

1. Vers une interface graphique interactive

Au niveau le plus bas de l'analyse de données, on trouve la description de la distribution d'une variable. Nous avons choisi cet exemple très simple pour montrer à quel point l'association de la graphique et d'un SIG peut transformer une pratique banale. Prenons l'exemple du logiciel Mapinfo : l'ouverture de la statistique sur une variable amène à l'écran l'affichage de la liste classique des indicateurs univariés. L'utilisateur peut donc prendre connaissance des valeurs, et continuer son travail. Il est incapable de décrire la forme de la série (bimodalité, asymétrie...) ou d'évoquer l'allure de la distribution géographique de sa variable. Dans un module statistique intégrant la graphique, la prise en compte de ce type d'information est conçue de manière visuelle et interactive. On affiche donc un diagramme de distribution de la variable (histogramme de fréquences), la moyenne est repérée par un trait vertical, l'écart-type par un trait horizontal (fig. 1). L'utilisateur est donc capable d'avoir une image précise de la distribution statistique de la variable qu'il étudie. De plus, il a accès à la géographie de ses données. En effet un clic sur une des barres de l'histogramme des fréquences sélectionne les objets spatiaux dans la carte, visualisant ainsi les individus qui composent une classe. Il en est de même pour la moyenne : en cliquant à droite de la moyenne, tous les individus qui lui sont supérieurs seront sélectionnés. En cliquant sur la barre représentant l'écart-type, ce sont les individus situés dans un intervalle d'un écart-type de part et d'autre de la moyenne qui seront sélectionnés.

Figure 1 : Diagramme de distribution d'une variable

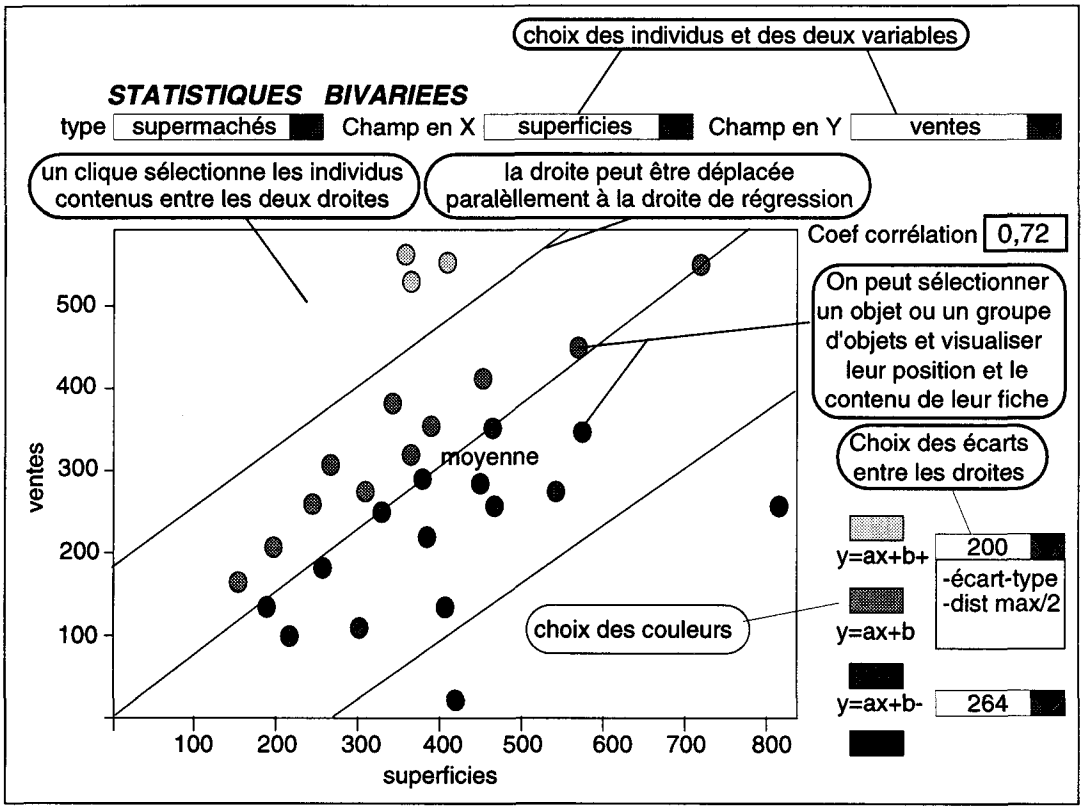


Ce type d'interface, dont le fonctionnement est basé sur l'image, image de la distribution statistique, image de la distribution géographique, permet à l'utilisateur de mener une réflexion géographique sur ses données. Il s'agit donc bien d'utiliser les méthodes de la graphique, rendues familières aux géographes par l'entremise de la cartographie, pour enrichir les outils existants de nouvelles fonctionnalités et pratiques.

Les mêmes principes de visualisation et d'interactivité sont appliqués aux modules d'analyses bivariées et multivariées. Prenons l'exemple d'une régression simple entre deux variables. L'utilisateur voit s'afficher le nuage de

points et la droite de régression est automatiquement tracée. De plus, deux droites sont tracées de part et d'autre de la droite de régression, à une distance d'un écart-type ou à une distance fixée par l'utilisateur (fig. 2). Les points situés dans les espaces séparés par les droites sont affectés d'un symbole qui sera le symbole des objets dans la carte. On a donc un moyen de constituer des classes selon la méthode classique de cartographie des résidus d'une régression. Mais les fonctionnalités ne s'arrêtent pas là. En effet, il y a interaction entre la carte et le graphique de la régression. On peut donc cliquer sur un point et le voir sur la carte, sélectionner un groupe d'objets sur la carte (par exemple les communes situées à 15 kilomètres d'un centre) et les visualiser sur le graphique. De même, un double-clic sur un point entraîne l'ouverture de la fiche de l'objet et permet de prendre connaissance de l'ensemble des données attachées à l'objet.

Figure 2 : Nuages de points et droite de régression



2. Création d'espaces non-topographiques

L'espace du graphique de la régression doit être conçu comme un espace cartographique. Un travail sur les villes françaises a donné lieu à la création d'une base contenant des objets du type villes. Il existe dans le logiciel un générateur de fonctions mathématiques permettant de combiner les valeurs de plusieurs variables à l'aide d'opérateurs (plus, moins, log...). Les résultats du calcul sont inscrits dans la fiche de chaque ville à l'aide de l'outil de remplissage qui peut utiliser la fonction mathématique pré-définie pour calculer pour chaque objet la valeur souhaitée. Il a donc été simple de créer deux champs dans la fiche de l'objet et de les remplir avec le log de la population et le log du rang. Leur visualisation sous forme de nuage de points permet d'obtenir un graphique rang-taille classique. On peut alors cartographier l'écart des points représentant les villes à la droite ajustant le nuage, mais aussi réaliser une cartographie dans le nuage de points. On peut, par exemple, visualiser la proportion de cadres dans la population active sur le graphique rang taille, et comprendre immédiatement le lien entre taille des villes et proportion de cadres. Le graphique cartésien est approché comme un espace, il est donc naturel de lui appliquer des méthodes de représentation cartographique. Le même type d'approche est appliqué aux analyses factorielles : les plans factoriels sont alors conçus comme des espaces cartographiables.

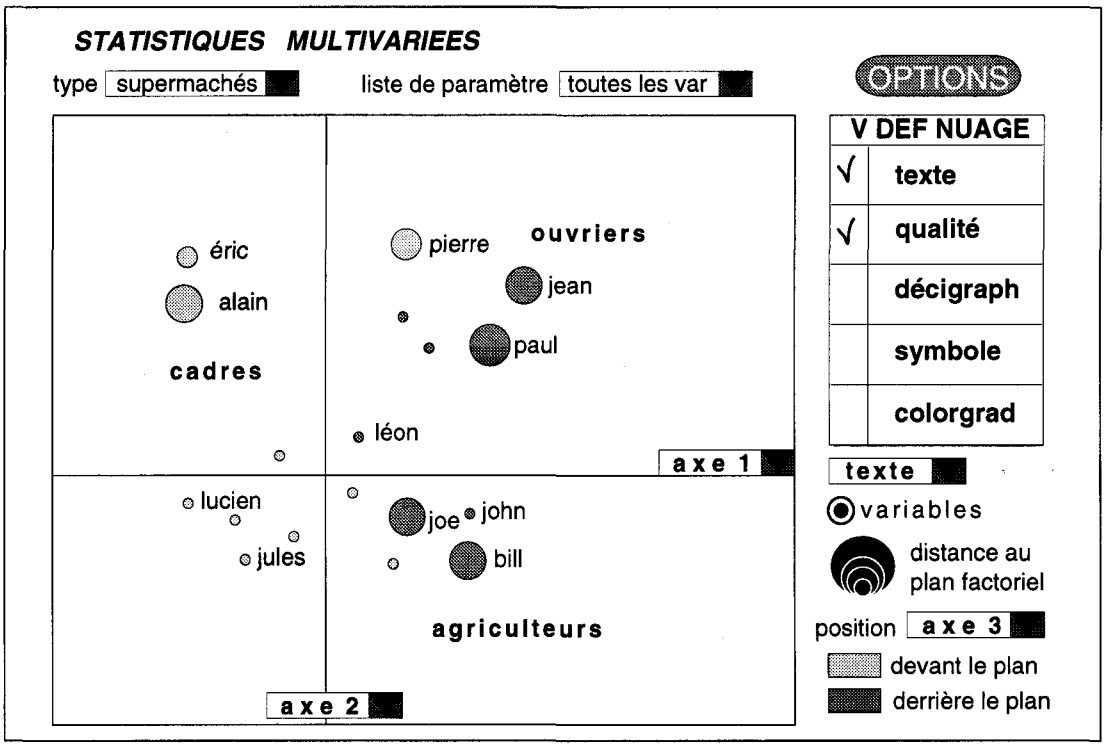
3. Graphique et méthodologie d'analyse statistique

Ce type d'outils devrait amener à de nouvelles pratiques notamment dans le domaine de l'analyse exploratoire. Il existe déjà de nombreux outils de statistiques exploratoires, intégrant des fonctionnalités avancées de visualisation. Il est courant de pouvoir visualiser des plans factoriels en trois dimensions, avec la possibilité de choisir un point de vue et de pénétrer dans le nuage de points. L'optique adoptée pour le développement de ce module se différencie par les modes de visualisation choisis mais plus encore par l'intégration de la composante spatiale des données. En cela leur élaboration peut être réalisée valablement par des géographes, car il s'agit bien d'intégrer les savoir-faire de l'analyse spatiale tout en créant de nouvelles pratiques associées.

Chaque étape de la démarche d'analyse va se trouver modifiée. La sélection des données s'enrichit de critères géographiques (temps de parcours entre deux lieux, inclusion dans une surface, proximité d'un objet). La sélection des individus se réalise également avec un filtre logique ou à l'aide de la souris, sur la carte affichée à l'écran. Ces outils de sélection, courants dans les SIG, rendent économiquement viable la multiplication de lancement d'analyses utilisant des filtres géographiques variés. La réflexion sur les données s'en trouve qualitativement augmentée.

La prise en compte des résultats est, elle aussi, modifiée ; en effet au lieu d'avoir à explorer un tableau d'indicateurs de qualité pour appréhender la validité de ses résultats, l'utilisateur en prendra connaissance graphiquement. Dans le cas d'un nuage factoriel, on pourra cartographier la distance au plan par un cercle de taille variable pour chaque ville. La position du point, devant ou derrière le plan par rapport au troisième axe, sera représentée par une couleur. L'utilisateur pourra choisir de voir la valeur sous forme de texte, il lui suffira alors d'ajouter « une manière de voir » ses objets villes sous une forme texte, en affichant la valeur de la distance au plan (fig. 3). Cette application de la puissance du langage graphique à la prise en compte de résultats d'analyses statistiques augmente considérablement l'ergonomie. De plus l'utilisateur se rend très vite compte de la validité de son analyse.

Figure 3 : Objets situés sur l'axe 3 et projetés sur le plan factoriel 1/2



A cet apport lié à la graphique, s'ajoute celui de l'interactivité avec la composante spatiale de la donnée. Chaque point du plan factoriel ou du nuage de régression est lié à sa représentation sur la carte. L'utilisateur a donc la possibilité de visualiser simultanément la position d'un objet au sein de l'espace créé par l'analyse et au sein de l'espace géographique.

Enfin, l'exploitation des résultats peut se faire sous forme d'une visualisation cartographique. L'analyse est conçue pour permettre de déboucher sur un document final qui en présente les résultats. Pour la régression chaque point est caractérisé par son écart à la droite. Des groupes de points sont créés entre chacune des droites. Il est alors possible d'appliquer la même manière de voir les objets dans le graphique et dans la carte. Si l'on a choisi de représenter l'appartenance à un groupe par une couleur, on visualisera les groupes selon leur couleur sur la carte.

4. Nécessité d'assister l'interprétation des résultats

Les travers à éviter dans l'emploi de ces méthodes d'analyses exploratoires sont nombreux. Le principal est d'utiliser une méthode de traitement inadaptée à la nature des données. L'expérience de la mise à disposition d'outils de cartographie à des utilisateurs non-cartographes, montre bien comment la méconnaissance des règles de base amène à des erreurs grossières. L'utilisation de méthodes statistiques suppose une attention constante à la nature des données traitées, ainsi qu'à la qualité des résultats obtenus. L'aspect « scientifique » de ces méthodes amène un utilisateur non averti à une attitude de confiance vis-à-vis des résultats de l'analyse. Le risque est alors grand de déboucher sur une analyse exhaustive des résultats perdant de vue la puissance synthétique offerte par ce type d'outil. Ce risque se double de la nécessité de posséder un bon niveau de connaissance, tant dans le domaine de l'interprétation statistique que de l'analyse géographique, ainsi que dans la prise en compte des résultats. Le logiciel doit donc permettre de visualiser le résultat global de l'analyse et de descendre au niveau des individus en proposant des outils efficaces de mesure de la pertinence de l'analyse. Si cet objectif n'est pas réalisé, l'utilisateur diluera l'information statistique apportée par les méthodes d'analyse statistique, dans l'observation de phénomènes isolés et peu significatifs. A cet empirisme s'ajoute un autre danger de ce type d'outil, amené par la possibilité d'adapter un modèle aux données. La tentation est alors grande pour l'utilisateur de généraliser son résultat. Ce type d'analyse ne fait que plaquer des modèles, qui n'ont un pouvoir explicatif que pour un exemple. De l'efficacité de l'interface dépend la possibilité de repérer un individu aberrant, un effet de taille ou encore une inversion de corrélation, et donc d'amener un regain de qualité dans l'interprétation. La possibilité de connaître facilement et visuellement les indices de qualité d'un individu permet d'envisager des effets de localisation que l'on n'aurait pas devinés autrement.

D'un point de vue informatique la mise au point des interfaces nécessite un SIG permettant de jongler entre les différents espaces et les modes de représentation. Les fonctionnalités d'une base de données orientée objets, associées aux outils de visualisation, rendent possible d'un point de vue technique la réalisation de ce type d'outil. Ainsi dans MacMap, on peut voir un objet selon plusieurs méthodes de manière simultanée. On peut visualiser l'objet ville selon un champ de texte, un point de taille proportionnelle à un champ numérique, une couleur représentant une catégorie. Ces possibilités de visualisation n'ont pour limites que les capacités du système oculaire ; en effet la multiplication des méthodes de visualisation crée rapidement une image confuse. Le rôle du logiciel est donc de proposer un choix de méthodes de visualisation optimisées limitant l'apparition d'images confuses. De même, c'est cette possibilité de voir un même objet de différentes manières au sein d'espaces différents, sans pour autant le dupliquer, qui rend possible le développement informatique de tels outils. On ne raisonne plus sur un objet repéré par ses coordonnées géographiques, mais bien sur un objet dont n'importe quelle donnée peut être utilisée pour le repérer dans un espace géographique ou non.

L'objectif est donc de mettre au point des outils créant de nouveaux systèmes de coordonnées. D'un histogramme de fréquences, à un plan factoriel en passant par un nuage de points, on est à chaque fois en présence d'un de ces espaces recréés. L'application à ces espaces de méthodes de visualisation cartographique, et leur mise en interaction devraient offrir aux géographes de puissants outils d'analyse.

Bibliographie

[1] BERTIN J. : *Sémiologie graphique,* Paris, Mouton-Gauthier-Villars, 1967, 431 pages

[2] DYKES J.A. : « Pushing maps past their established limits : A unified approach to cartographics visualization », *Proceedings of GIS research,* 1995, pp. 78-85

[3] MacEACHERN & TAYLOR D.R. Fraser : *Visualisation in modern cartography,* New-York, Pergamon/Elsevier, 1994, 368 pages

[4] de GOLBÉRY L. & ORHAN J.-M. : « Sémiologie graphique le retour ? », Barcelone, *Actes de la 17e conférence internationale de cartographie,* 1995, pp. 1969-1975