LA MESURE D'INDICATEURS LOCAUX CONSIDÉRÉE COMME UN "BOOTSTRAP SPATIAL"

Jean-Christophe FOLTÊTE
THEMA UPRESA 6049
Université de Franche-Comté

Résumé

Notre démarche consiste à lier deux techniques sans rapport apparent : bootstrap et indicateurs locaux. Le bootstrap est une méthode de rééchantillonnage destinée à observer la stabilité de toute mesure statistique ; les calculs d'indicateurs locaux consistent à itérer un calcul de lien entre variables en sélectionnant successivement le voisinage spatial des unités étudiées. Dans les deux cas, l'analyse ne se fonde plus sur un modèle unique mais sur une grande série de sous-échantillons. Sous cet angle, les indicateurs locaux servent à observer la stabilité spatiale de toute mesure statistique. Nous tentons de montrer leur intérêt par rapport à un modèle global, à travers un exemple précis : la relation entre densité de population et mobilité quotidienne des actifs en Franche-Comté.

Mots-Clés

Bootstrap, fenêtre coulissante, filtrage spatial, indicateur local, modèle global, modèle local, pondération gaussienne, voisinage

En analyse spatiale, il est souvent question d'adapter des méthodes statistiques aux données géographiques qui ont la particularité d'être structurées par des rapports topologiques. Nous suivons ici une démarche un peu différente, en assimilant une technique explicitement spatiale (les indicateurs locaux) à une forme particulière de bootstrap, principe statistique général. Notre objectif est de souligner l'intérêt et la spécificité des filtrages spatiaux mono-, bi- ou multivariés.

Bootstrap et indicateurs locaux n'ont pas de rapport apparent : ces techniques ont été conçues dans des contextes différents, c'est pourquoi nous commençons par les présenter de façon distincte, avant de recenser leurs points communs. La prise en compte des rapports topologiques amène une discussion autour du choix de la fonction de décroissance appliquée aux distances. Enfin, à travers la question du rapport entre densité et mobilité des actifs en Franche-Comté, nous illustrons l'ensemble du propos par un exemple concret.

1. Le bootstrap : un contrôle de la stabilité des formes statistiques

Tout échantillon d'apprentissage conditionne le modèle statistique qu'il permet de créer. Cette relation biaise le caractère généralisable des résultats, ce qui explique les récentes questions autour des notions de stabilité et de robustesse. Les méthodes de rééchantillonnage (validation croisée, jacknife, bootstrap) ont été développées dans le but explicite de réduire le biais lié à l'apprentissage d'un modèle. Parmi celles-ci, le bootstrap s'avère des plus efficaces [5] [6].

Son principe est simple (fig. 1). Admettons un modèle statistique construit à partir de n individus ; il consiste à simuler un grand nombre de nouveaux échantillons, en pratiquant un tirage avec remise dans l'échantillon initial. La taille de ces sous-échantillons, inférieure ou égale à n, doit être constante.

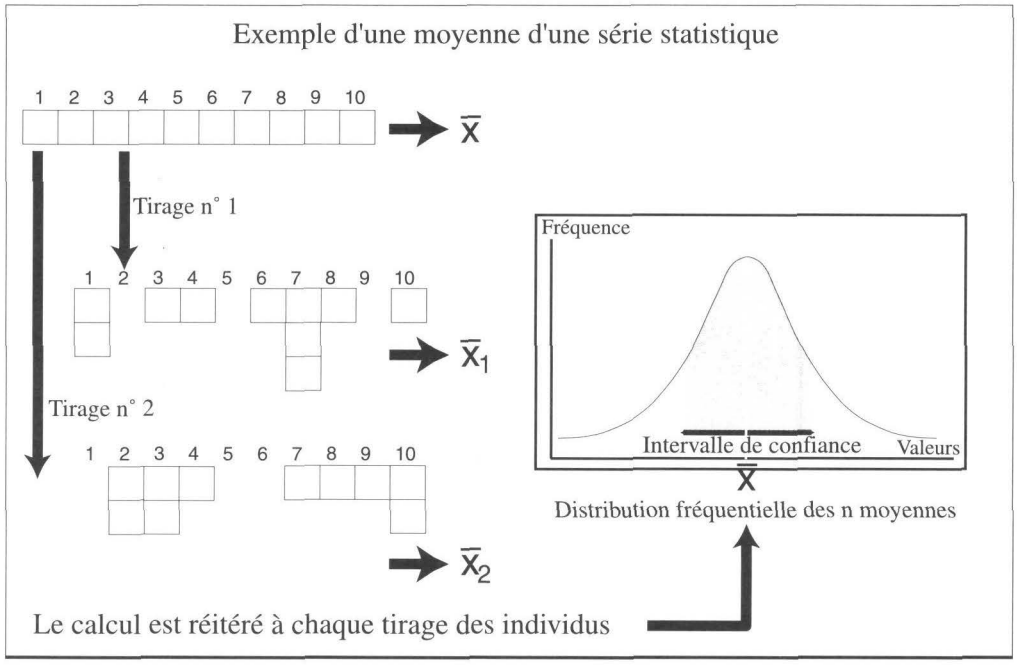

Figure 1 - Exemple d'un tirage bootstrap pour le calcul d'une moyenne arithmétique

À chaque tirage, les paramètres du modèle sont calculés; au bout d'un certain nombre de tirages (si possible très supérieur à 30 et fonction des possibilités informatiques), on peut observer la variance de ces paramètres. Si celle-ci est faible, le modèle est jugé stable puisqu'il n'est pas sensible à des variations d'échantillonnage dont la distribution n'est pas spécifiée. En observant la distribution des paramètres, on peut leur définir un intervalle de confiance, sans l'hypothèse d'une distribution particulière. Non paramétrique, le principe du bootstrap s'adapte à n'importe quel modèle (régression, analyse de variance, analyses factorielles, etc.); toutefois, il n'est pas encore beaucoup utilisé en analyse spatiale.

2. La mesure locale : une analyse par "zooms" successifs

Les indicateurs locaux relèvent des calculs de voisinage très répandus en analyse d'image [4]. Dans ce domaine, il est courant d'utiliser des opérateurs focaux (ou filtres) pour déterminer un contexte spatial autour de chaque point. Les fenêtres coulissantes servent le plus souvent au calcul de texture ou d'arrangement spatial, c'est-à-dire à une mesure monovariée. La forme de ces fenêtres peut être variable : carrée, circulaire ou quelconque [11].

L'adaptation de ces fenêtres à des mesures bi-variées ou multivariées est peu courante, malgré un principe très simple (fig. 2). Des recherches récentes présentent ce principe, en premier lieu sur des unités spatiales régulières : carroyage ou image numérique [2] [3]. D'autres travaux montrent la pertinence de cette approche appliquée à des unités irrégulières (découpages administratifs), sous la forme d'*indicateurs locaux d'association spatiale* [1]. Nous avons par ailleurs adopté ces principes en utilisant l'expression d'*ajustement local,* pour étudier les liens entre une partition en types morphologiques et une série de données sociales [7] à une échelle communale.

Les mesures locales s'avèrent très intéressantes en analyse spatiale, puisqu'elles permettent de lier le calcul statistique aux différentes configurations de l'espace. Précisons par ailleurs qu'une telle approche n'est pas équivalente à l'analyse locale présentée par L. Lebart [8]. En effet, selon cet auteur, les calculs de variance et de corrélation sont modifiés pour ne concerner que les formes locales au sens d'une certaine contiguïté, mais les résultats gardent un aspect global. De la même façon, les analyses factorielles "locales" [9] [10] consistent à modifier les distances statistiques en introduisant les distances spatiales, sans obéir au principe de mesures itérées localement.

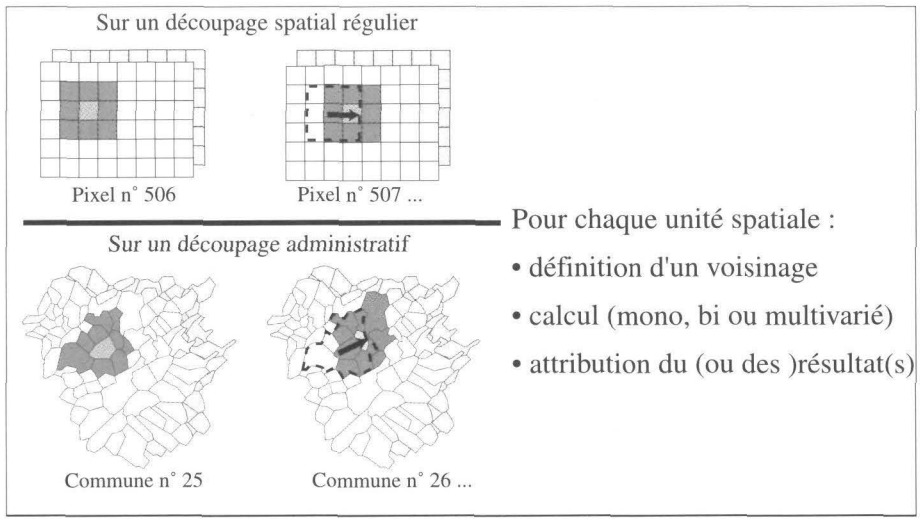

Figure 2 - Principe général des indicateurs locaux

3. Bootstrap et indicateurs locaux : deux types de pondération des unités spatiales

Dans la mesure où le tirage des nouveaux échantillons est effectué avec remise, on peut considérer un échantillon-bootstrap comme une pondération aléatoire des individus, en fixant la somme des poids égale à la taille de l'échantillon (fig. 3). Dans chaque nouvel échantillon, certains individus ont donc une influence accrue, au détriment d'autres qui ne participent plus au calcul.

L'ajustement local est une mesure itérative : les individus sont tour à tour associés à un certain voisinage dans lequel on effectue le calcul voulu. Chaque itération peut donc être considérée comme une pondération binaire, où le poids des individus est unitaire à l'intérieur de la fenêtre et nul à l'extérieur (fig. 3). On passe ainsi d'une analyse globale à une série d'analyses locales, où la taille des échantillons (inférieur à n) est fonction de la fenêtre de voisinage. C'est donc par un changement d'échelle qu'on peut évaluer localement la valeur d'une mesure réalisée en principe sur l'ensemble des individus.

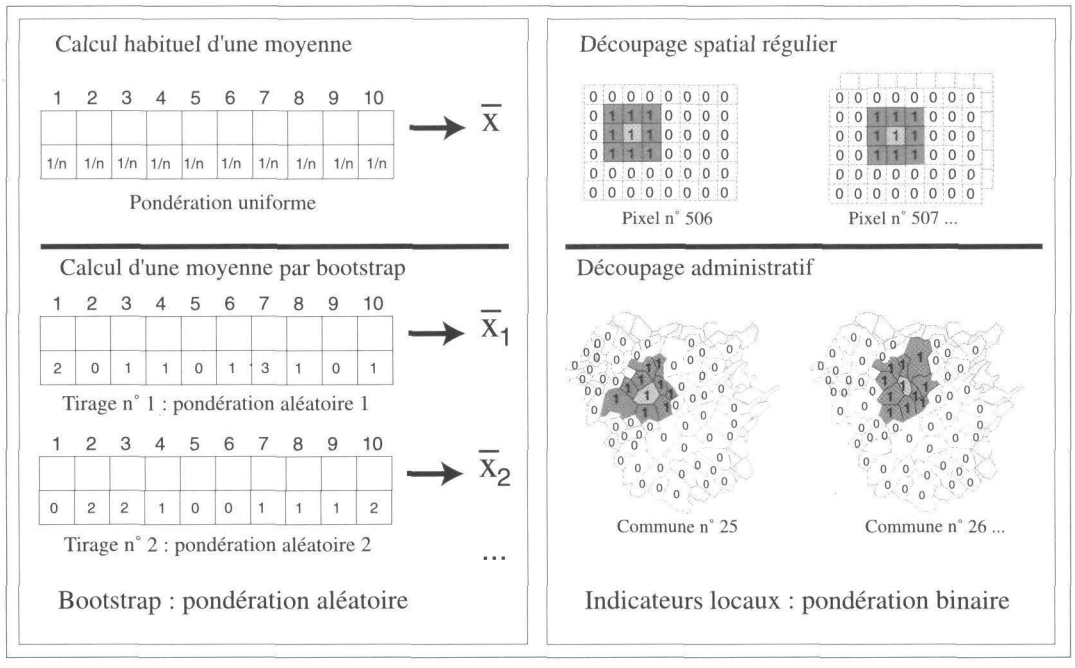

Figure 3 - Bootstrap et indicateurs locaux comme des pondérations particulières

Bootstrap et mesures locales sont des techniques différentes qui ont deux points communs essentiels : dans les deux cas, il s'agit du calcul itératif d'un certain modèle statistique ; à chaque itération les individus sont pondérés de façon spécifique. Sous cet angle, nous allons tenter de concilier les avantages de chacune : il s'agit de contrôler la stabilité des paramètres d'une analyse tout en prenant en compte la structure topologique des unités spatiales.

Un indicateur local peut donc être considéré comme une variante du bootsrap, adaptée à l'analyse spatiale. Celle-ci consiste à construire n nouveaux échantillons, où les individus sont pondérés non de façon aléatoire mais en fonction de leur agencement. Nous partons du principe que la production "aléatoire" des échantillons bootstrap (comme d'ailleurs celle de tout échantillon en statistique classique) n'est pas une condition toujours intéressante à respecter en analyse spatiale, dans la mesure où, en général, on connaît a priori la configuration des individus. Alors que le bootstrap revient à ignorer toute structure a priori, il peut être avantageux d'inclure celle-ci dans le calcul pour réaliser une analyse en fonction de l'espace. D'autre part, l'ajustement local est ici considéré comme un moyen efficace d'observer les variations spatiales des phénomènes étudiés.

4. Quelle fonction de la distance pour une influence décroissante des unités voisines ?

Pour aborder la définition d'un voisinage plus nuancé, nous substituons à la fonction de pondération binaire des fenêtres classiques une fonction continue fondée sur les distances entre unités spatiales. En conséquence, le problème du choix d'une taille de voisinage laisse place à celui de la fonction à appliquer aux distances. En modélisation gravitaire, l'exposant négatif ($y = d^{-\beta}$) vient intuitivement à l'esprit pour formaliser la décroissance des interactions avec l'éloignement. Ce choix peut aussi porter sur d'autres fonctions [2] ; ainsi, pour un certain nombre de raisons (forme logistique de la courbe de pondération, pas de problème pour la pondération centrale avec une distance nulle), il nous semble pertinent d'utiliser une fonction gaussienne de décroissance, de la forme :

$$y = exp(-\beta.d^2).$$

Dans la formule précédente, le coefficient β remplit un double rôle :

– sa valeur définit l'étendue spatiale du voisinage : si elle est faible, les pondérations sont très proches d'une unité à ses voisines, ce qui induit un fort lissage ; si au contraire elle est forte, le rôle de l'unité centrale est prépondérant ;

– il permet aussi l'utilisation de toutes sortes de distances, sans faire jouer les unités (distances physiques, de contiguïté, temporelles).

Le choix de ce paramètre peut être instrumenté, par maximisation d'un certain critère statistique, comme les moindres carrés [2]. Toutefois, nous ne sommes pas persuadés que la valeur optimale, d'un point de vue numérique, corresponde à celle qui donne lieu aux résultats les plus intéressants pour l'interprétation. En effet, dans l'exemple qui suit, la valeur optimale de ce paramètre est très forte et conduit à une production de cartes très pointillistes. C'est pourquoi nous en restons ici à une définition empirique du paramètre β, par comparaison des cartes de résultats.

5. Exemple d'application : densité de population et mobilité des actifs en Franche-Comté

Pour illustrer la technique présentée, nous proposons d'étudier la relation entre la densité de population et la part d'actifs travaillant hors de leur commune de domicile, sur les 1 786 communes franc-comtoises. Les deux variables sont transformées (respectivement par les fonctions logarithme et exponentielle) afin de travailler sur des formes de distribution à peu près normales.

Densité et mobilité des actifs sont a priori corrélées positivement, dans la mesure où les migrations travail-domicile sont surtout le fait des agglomérations urbaines. Pourtant, les deux phénomènes

n'apparaissent pas calqués l'un sur l'autre : les principaux centres urbains (Besançon, Belfort, Dole et Vesoul) connaissent une forte densité, mais leurs actifs sont en grande partie sédentaires, en comparaison des communes périphériques. Cette dissociation visible sur le graphe de régression contribue à une faible corrélation ($r = 0,33$).

5.1. Variation spatiale de la corrélation

Faiblement positive, la corrélation entre les deux phénomènes étudiés est difficile à cerner sur l'ensemble de la région. Nous appliquons donc un calcul de corrélations locales en pondérant les voisins de chaque commune par une fonction gaussienne (distances physiques en kilomètres, paramètre $\beta = 0,1$). Par comparaison avec le coefficient global ($r = 0,33$), la mesure de la relation varie à présent de –0,73 à +0,68 : localement, mobilité des actifs et densité ne sont pas toujours liés positivement.

La figure 4 montre la distribution spatiale du coefficient de corrélation. Les agglomérations urbaines se détachent avec des coefficients négatifs, c'est-à-dire avec une proportionnalité inverse entre mobilité et densité : l'accent est donc mis sur des structures de type centre-périphérie, où une couronne aux actifs très mobiles gravite autour d'un centre plus sédentaire du point de vue de l'emploi. Toutefois, les valeurs positives

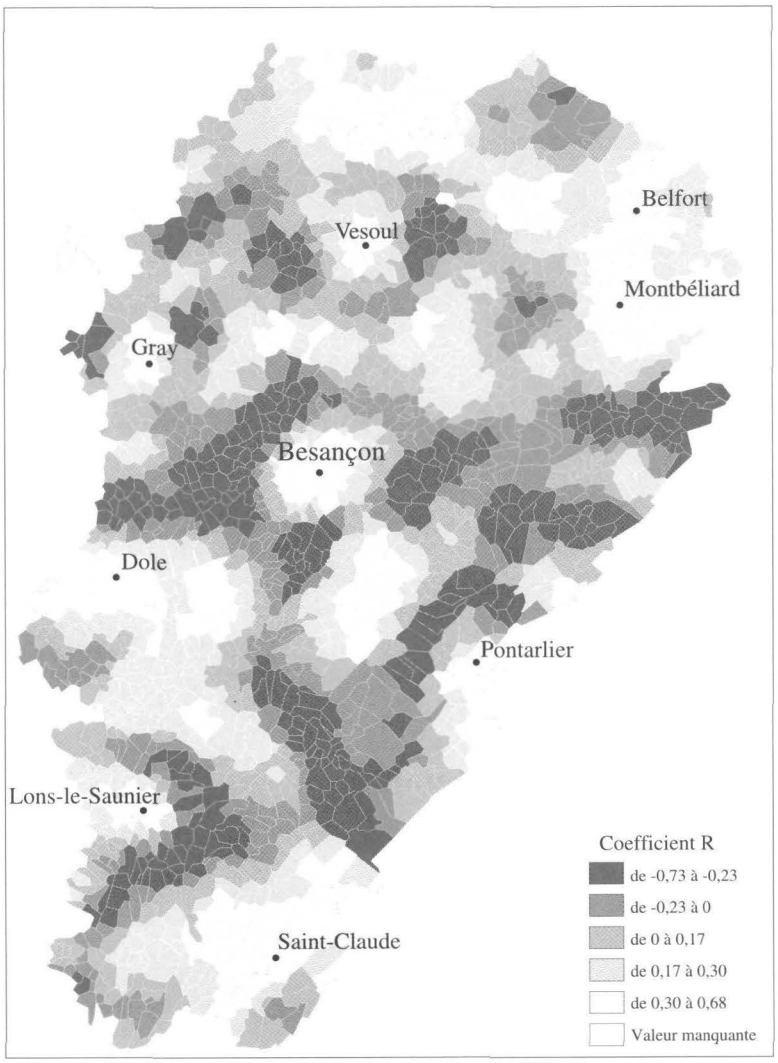

Figure 4 - Distribution spatiale du coefficient de corrélation entre densité et mobilité

s'étendent bien au-delà des grandes zones urbaines, comme dans les Vosges sâonoises ou sur les premiers plateaux jurassiens (au sud-est de Besançon). Il s'agit là probablement d'une attraction exercée en milieu rural par des pôles secondaires, entourés d'une campagne très peu peuplée.

D'autre part, des zones de forme linéaire, qui délimitent souvent un périmètre autour des agglomérations précédentes, se distinguent par une relation positive : dans les communes les plus peuplées se trouvent davantage de personnes qui travaillent à l'extérieur (a priori dans les agglomérations). La zone de Belfort-Montbéliard se distingue avec un coefficient toujours négatif ; il s'agit là peut-être d'une mobilité particulière induite par l'industrie automobile et le caractère multipolaire de la structure urbaine : il est probable qu'à cet endroit, les communes les moins peuplées comportent peu d'activités et engendrent des flux plus importants, à destination des lieux de production.

5.2. Synthèse des relations locales

Les variations spatiales de la relation entre densité et mobilité des actifs obéissent à une logique qui s'inscrit dans les rapports ville/campagne. Toutefois, au vu du seul coefficient de corrélation, il est difficile de tirer des conclusions simples : localement, nous ne pouvons pas déterminer si une relation négative (ou

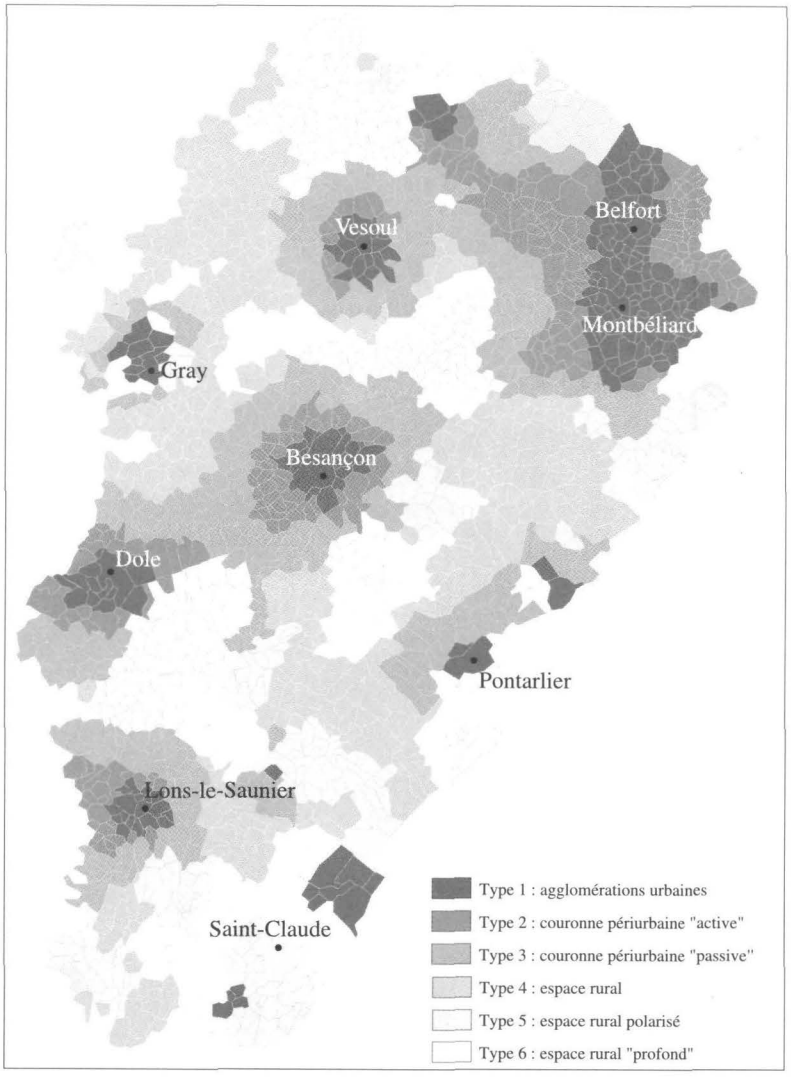

Figure 5 - Typologie des relations locales entre densité et mobilité des actifs

positive) représente des fortes valeurs de densité, si ces valeurs sont homogènes, etc. Il faudrait pour cela associer à la carte des coefficients celles des paramètres qui ont contribués à leur calcul ; l'étude locale de la relation entre deux phénomènes nécessite donc l'examen de cinq documents : moyennes et variances locales respectives de la densité et de la mobilité, coefficient de corrélation linéaire. Dans un souci de simplification, nous choisissons ici de réduire ces cinq cartes en une typologie, en appliquant une classification par la méthode des moyennes mobiles.

Les 6 types (fig. 5) issus de cette classification sont les suivants :

- **type 1 : agglomérations urbaines**

Les densités sont fortes mais très variables, du fait de la présence de centres urbains. En revanche les actifs sont dans l'ensemble très mobiles. La corrélation négative précise que les communes périurbaines, moins densément peuplées, sont à l'origine de mouvements journaliers plus intenses.

- **type 2 : couronne périurbaine "active"**

La densité et la mobilité des actifs sont à la fois importantes et homogènes. La relation toujours négative montre une plus grande sédentarité des unités les plus peuplées : l'emploi local maintient donc une partie des actifs.

- **type 3 : couronne périurbaine "passive"**

Logiquement, les niveaux de densité et de mobilité décroissent à mesure qu'on s'éloigne des centres. En revanche, la corrélation s'inverse ici avec un coefficient positif. Ce sont donc les communes les plus importantes de la couronne qui alimentent les flux d'actifs, ce que nous pouvons interpréter comme une plus grande dépendance économique.

- **types 4 et 5 : espace rural polarisé**

Nous constatons à nouveau une décroissance de la densité et de la mobilité. Dans la mesure où les communes les moins peuplées contiennent les actifs les plus mobiles, il s'agit ici d'espaces ruraux sous l'influence soit de pôles secondaires locaux, soit de villes relativement lointaines. Le type 5 se distingue du type 4 par une grande variabilité de la mobilité, qui montre une polarisation locale plus importante. Les zones correspondantes sont par exemple celles de Poligny, Champagnole, Le Valdahon ou Saint-Claude.

- **type 6 : espace rural "profond"**

De même que nous trouvons les plus faibles densités de Franche-Comté, les actifs sont en majeure partie sédentaires. Localement, ces deux phénomènes varient peu. La corrélation positive indique que les foyers de peuplement sont sous l'influence de pôles extérieurs.

La typologie décrite ne correspond pas seulement à une partition en classes de composition homogène, elle intègre la relation (statistique) des phénomènes étudiés. Ainsi densité et mobilité ne sont-elles pas liées de façon uniforme dans l'espace : la mesure locale de cette relation renseigne sur la structure spatiale, qu'elle soit polarisée par un centre important ou un bourg rural, ou au contraire homogène.

Conclusion

Au contrôle de la stabilité statistique du bootstrap "classique", par ailleurs très utile, nous préférons celui la mesure de la stabilité spatiale. La cartographie des variations d'une relation permet d'aborder une étude par le caractère résolument spatial des données géographiques, plutôt que de raisonner directement en termes statistiques. S'agissant de pratiquer des moyennes locales, c'est-à-dire de faire un lissage spatial, la technique du voisinage est maintenant largement utilisée. Quant à prendre en compte simultanément des mesures de relations, nous touchons là une démarche encore inhabituelle.

Le calcul d'indicateurs locaux peut s'inscrire dans plusieurs approches. Ainsi cette démarche peut être utilisée pour la sélection de zones à l'intérieur d'un espace important, si on veut analyser un phénomène dont on ne connaît pas la délimitation pertinente (elle peut ainsi guider un échantillonnage…). La même démarche

sert aussi à enrichir l'analyse "classique" : l'application présentée montre en effet qu'une simple corrélation calculée sur une vaste série d'unités spatiales masque un grand nombre de configurations locales, qui sont parfois contradictoires. De plus, les cartes produites présentent une forte autocorrélation spatiale (dépendant du choix de la fonction de distance), ce qui permet de construire des typologies aux zonages bien définis.

Comme le bootstrap, il ne s'agit ni d'une méthode précise, ni d'une alternative à toute autre démarche : c'est un principe qui s'adapte à beaucoup de questions différentes sans grande contrainte ; il est possible par exemple d'étudier les distributions spatiales de rapports r^2 d'analyse de variance [7], ou tout autre coefficient. L'extension à des données multivariées est techniquement possible, mais pose des problèmes de gestion de résultats, puisque le nombre d'indicateurs locaux augmente plus vite que celui des variables ! Un deuxième niveau de traitement est alors nécessaire pour synthétiser les résultats (comme la classification opérée ici), mais la démarche devient nettement plus lourde.

Références bibliographiques

[1] ANSELIN L., 1995 : Local Indicators of Spatial Association—LISA, *Geographical Analysis*, vol. 27, n° 2, pp. 93-115

[2] BRUNSDON C., FOTHERINGHAM A.S., CHARLTON M.E., 1996 : Geographically Weighted Regression : A Method for Exploring Spatial Nonstationnarity, *Geographical Analysis*, vol. 28, n° 4, pp. 281-298

[3] CHARLTON M., FOTHERINGHAM A.S., BRANSDON C., 1996 : The Geography of Relationships : an Investigation of Spatial Non-Stationarity, in *Spatial Analysis of Biodemographic Data*, John Libbey Eurotext, Paris, pp. 23-47

[4] COCQUEREZ J.-P., PHILLIPS S. (coordinateurs), 1995 : *Analyse d'image : filtrage et segmentation*, Masson, Paris, 457 pages

[5] EFRON B., 1979 : Bootstrap methods : another look at the Jacknife, *Annals of Stat.*, vol. 7, pp. 1-26

[6] EFRON B., TIBSHIRANI R.J., 1993 : *An Introduction to the Bootstrap*, Chapman and Hall, New York

[7] FOLTETE J.-C., 1998 : *Production sociale et dimension visible du paysage ; analyse géographique*, Thèse de doctorat, Université de Franche-Comté, 380 pages

[8] LEBART L., 1969 : L'analyse statistique de la contiguïté, *Publications de l'ISUP*, vol. 28, pp. 81-112

[9] LE FOLL Y., 1982 : Pondération des distances en analyse factorielle, *Statistiques et Analyse des Données*, n° 7, pp. 13-31

[10] THIOULOUSE J., CHESSEL D., CHAMPELY S., 1995 : Multivariate Analysis of Spatial Pattern : a Unified Approach to Local and Global Structure, *Environmental and Ecological Statistics*, n° 2, pp. 1-14

[11] TOMLIN C.D., 1990 : *Geographic Information Systems and Cartographic Modelling*, Prentice-Hall, London, 249 pages