Agrégation statistique et sémantique : un opérateur multi-échelles

Laurent RAYNAL,
IGN/COGIT,2, Avenue Pasteur 94160 Saint-Mandé (France),
tel : + 33 - 1 - 43 - 98 - 85 - 44, fax : + 33 - 1 - 43 - 98 - 81 - 71
Pierre DUMOLARD,
LAMA, Espace Serge Martin,BP 53X,38041 Grenoble Cedex (France)
tel : + 33 - 76 - 51 - 45 - 96

De par la diversité des sources de saisie, les données géographiques n'ont pas la même résolution et se rapportent donc à des échelles différentes. Or gérer des données à différentes échelles est nécessaire en géographie ou en géo-statistique car ces disciplines doivent maintenir, ou créer plusieurs niveaux de synthèse pour l'information. De plus, la gestion de telles données suscite une formulation différente suivant les disciplines : en géographie, on parlera de niveaux d'organisation spatiale ; en informatique, de relations hiérarchiques ; en statistique de hiérarchie de partitions.

Une convergence apparaît pourtant en termes d'outils et de méthodes : ce sont les mêmes données qui sont manipulées. Aussi, l'écueil actuel pour mettre en commun ces différents savoirs réside dans la communication entre acteurs de chaque discipline, dans l'organisation d'un message compréhensible par tous. Le groupe de l'axe B Multi-échelles du PSIG (Programme de recherche en Sciences de l'Information Géographique) s'est constitué pour atteindre de tels objectifs et cet article a pour but de présenter un état d'avancement de ses travaux.

Il s'agit donc de manipuler des données appartenant à différents niveaux d'organisation de l'espace. Or le transfert d'un niveau d'organisation à l'autre dépend des niveaux de départ et d'arrivée, dépend de la définition de chacun des niveaux (est-ce un maillage régulier de l'espace, est-ce un ensemble de zones construites par interpolation, est-ce un ensemble de zones délimitées arbitrairement ?). De toute évidence, la complexité du problème nécessite de poser un certain nombre d'axiomes, afin de définir le contexte de notre étude.

En premier lieu, on se placera dans un espace de géométrie euclidienne (définition de la notion de distance). Puis on considérera une partition de l'espace sous la forme de polygones disjoints, ni chevauchants, ni inclus. Enfin, chaque polygone porte des attributs de nature quantitative (population, surface utile...) et/ou des attributs de nature qualitative recodés numériquement (1 <-> ...). Ainsi, des polygones plus grands peuvent être créés par jonction de polygones élémentaires contigus. Une hiérarchie arborescente d'entités aréales emboîtées peut être définie et manipulée. Alors, des informations de même nature ou de nature différente peuvent être associées à chaque niveau de l'arborescence.

L'opérateur d'agrégation permet de transférer des informations d'un niveau inférieur vers un niveau supérieur. Etant donné que cette opération modifie la nature des informations (distribution spatiale..), des indices statistiques seront introduits afin de qualifier et quantifier de telles modifications de nature. En fait, un modèle statistique a été entièrement défini se fondant sur la notion de mesurabilité et évaluant la perte de mesurabilité dans l'opération d'agrégation. Dès lors, différents types d'agrégation sont apparus : l'agrégation géométrique (comme fusion de zones adjacentes) et l'agrégation sémantique dépendant de la nature des attributs (de type quantitatif absolu, de type quantitatif relatif, de type qualitatif ordonné, de type qualitatif nominal). Chacune de ces catégories fait alors l'objet d'une définition rigoureuse et est associée à une ou plusieurs opérations d'agrégation (moyenne, somme, moyenne pondérée, passage en fréquences). Enfin, une plate-forme d'expérimentation composée du logiciel statistique NewS et du SIG Smallworld permettra de valider ce modèle.