Éditorial Insee - guide de lecture

Vincent Loonis - *Responsable de la Division des Méthodes et Référentiels Géographiques (DMRG), Insee*
Marie-Pierre de Bellefon - *Responsable de la Section Méthodes d'Analyse Spatiale à la DMRG*

Pourquoi un nouveau manuel d'analyse spatiale ?

Cressie fut un des premiers à publier un "manuel de statistiques spatiales" (CRESSIE 1993a). Clair et détaillé, son ouvrage permet d'approfondir la théorie des statistiques spatiales. Il n'inclut cependant aucun guide d'utilisation pratique de ces méthodes. Depuis cette publication, les avancées théoriques et informatiques sont allées de pair avec l'accroissement de l'offre de données géolocalisées. De nombreux spécialistes ont à leur tour rédigé des manuels et autres guides de statistique spatiale : des très théoriquesPACE et al. 2009, GELFAND et al. 2010 ou ANSELIN 2013 aux guides d'utilisation du logiciel R : BIVAND et al. 2008, BRUNSDON et al. 2015 en passant par des ouvrages mêlant théorie et pratique comme HAINING 2003, SCHABENBERGER et al. 2004 ou FISCHER et al. 2009. Parmi les ouvrages francophones, ZANINETTI 2005 décrit la théorie de la statistique spatiale, tandis que CALOZ et al. 2011 s'intéressent à la géostatistique. Au sein même de l'Insee, Jean-Michel Floch a présenté en 2013 l'apport de la statistique spatiale pour l'étude des disparités socio-économiques (FLOCH 2013) et en 2015 ses réflexions sur la statistique spatiale en général.

L'objectif du manuel d'analyse spatiale est de répondre aux questions concrètes des chargés d'étude des instituts statistiques : que faire avec ces nouvelles sources de données géolocalisées ? Dans quels cas doit-on prendre en compte leur dimension spatiale ? Comment appliquer les méthodes de statistique et d'économétrie spatiale ? Contrairement aux manuels existants, la pédagogie est pensée spécifiquement en fonction des enjeux propres aux instituts statistiques : les exemples d'application utilisent des données collectées par la statistique publique et l'accent est mis sur la pratique et l'importance du choix des paramètres. Les fondements théoriques sont suffisamment approfondis pour permettre de comprendre les subtilités dans la mise en œuvre pratique des méthodes, tout en renvoyant aux ouvrages spécialisés les lecteurs désireux de connaître les extensions d'un niveau technique plus élevé. La majorité des chapitres présente des méthodes bien documentées et fréquemment utilisées, mais quelques-uns s'appuient sur des travaux innovants diffusés récemment. Parmi les thèmes abordés, le manuel Insee-Eurostat s'intéresse aux questions de sondage et de respect de la confidentialité ; autant de points importants pour les INS et très peu approfondis dans les ouvrages existants. Quelques chapitres ouvrent sur des notions peu utilisées actuellement à l'Insee comme la géostatistique.

Le panel d'auteurs mêle statisticiens de différents départements de l'Insee (Département de la Méthodologie Statistique, Département de l'Action Régionale, Département des Études et Synthèses Économiques) et professeurs universitaires (universités du Mans, Paris-Sud, Guyane, Agrosup et Inra Dijon). La rédaction du manuel a ainsi été l'occasion de favoriser les échanges entre le milieu de la statistique publique et le milieu académique.

Le plan du manuel

En 2008, le prix Nobel d'économie fut remis à Paul Krugman : le père de la nouvelle économie géographique. Cette récompense marque l'importance croissante de la prise en compte des

phénomènes spatiaux. Krugman décrit l'économie géographique comme "la branche des sciences économiques qui s'intéresse au lieu où les choses se produisent et aux relations entre elles" (KRUGMAN 1991). Cette citation illustre la démarche propre à toute étude d'analyse spatiale, quel qu'en soit le domaine d'application. L'analyste commence par décrire les lieux des observations, puis il mesure l'importance des interactions spatiales afin de pouvoir prendre en compte ces interactions grâce à un modèle pertinent. Ces trois étapes correspondent aux trois premières parties du manuel : Partie 1 : *Décrire les données géolocalisées* ; Partie 2 : *Mesurer l'importance des effets spatiaux* ; Partie 3 : *Prendre en compte les effets spatiaux*.

Le lieu est référencé dans un système d'information géographique grâce à ses coordonnées. Une des caractéristiques de l'analyse spatiale est donc que le support de l'observation, défini comme l'ensemble des coordonnées spatiales des objets à traiter, contient des informations potentiellement riches pour l'analyse. Pour les exploiter, le chargé d'étude commence le plus souvent par regrouper les données en fonction de leur proximité géographique. Il s'agit de la première étape avant d'explorer les caractéristiques de la localisation des données et de décrire l'évolution des variables dans l'espace. Ce regroupement est aussi un paramètre clé pour assurer le respect de la confidentialité des données diffusées par les instituts de statistique publique. Le premier chapitre du manuel : *Analyse spatiale descriptive* présente la façon dont on peut prendre en main les données avec le logiciel R et réaliser de premières cartes. Des notions de sémiologie cartographique sont également introduites. La deuxième étape d'une analyse spatiale est la définition du voisinage d'un objet. La définition du voisinage est indispensable pour mesurer la force des relations spatiales entre les objets, c'est-à-dire la façon dont les voisins s'influencent les uns les autres. L'enjeu du deuxième chapitre du manuel : *Codifier la structure de voisinage*, est de réussir à définir des relations de voisinage cohérentes avec les véritables interactions spatiales entre les objets. Ce chapitre présente plusieurs notions de voisinage, fondées sur la contiguïté ou sur les distances entre observations. La question du poids accordé à chaque voisin est aussi abordée.

Les données géolocalisées peuvent être réparties en trois catégories : données surfaciques, données ponctuelles et données continues. La différence fondamentale entre ces données n'est pas la taille de l'unité géographique considérée mais le processus générateur des données. Pour des données surfaciques, la localisation des observations est considérée comme fixe : c'est la valeur des observations qui suit un processus aléatoire. Par exemple, le PIB de chaque région est une donnée spatiale surfacique. Plus les valeurs des observations sont influencées par les valeurs des observations qui leur sont géographiquement proches, plus l'autocorrélation spatiale est élevée. Les indices d'autocorrélation spatiale permettent de mesurer la force des interactions spatiales entre les observations. Les versions globales et locales des indices d'autocorrélation spatiale sont présentées dans le chapitre 3 : *Indices d'autocorrélation spatiale*. Pour des données ponctuelles, la localisation des observations est la variable aléatoire. Il peut s'agir par exemple de la localisation des commerces au sein d'une ville. La force des interactions spatiales se mesure donc à l'aune de l'écart entre la distribution dans l'espace des observations et une distribution spatiale complètement aléatoire.Le chapitre 4 : *Les configurations de points* donne les méthodes et les outils permettant notamment de mettre en évidence les éventuelles attractions ou répulsions entre les différents types de points et la façon dont on évalue la significativité des résultats obtenus. Enfin, les données continues se caractérisent par le fait qu'il existe une valeur pour la variable d'intérêt en tout point du territoire étudié. En revanche ces données sont mesurées uniquement en un nombre discret de points. Il s'agit, par exemple, de la composition chimique du sol utile à l'industrie minière. Le chapitre 5 : *Géostatistique* présente les concepts fondamentaux permettant d'étudier les données continues : semi-variogramme, interpolation des données par les méthodes de krigeage,...

Les troisièmes et quatrièmes parties du manuel se concentrent sur l'étude des données surfaciques, auxquelles on a le plus souvent affaire dans les instituts de statistique publique. Parmi

les phénomènes spatiaux qui affectent les données surfaciques, on peut distinguer dépendance spatiale et hétérogénéité spatiale. La dépendance spatiale désigne une situation où la valeur d'une observation est liée aux valeurs des observations voisines (soit elles s'influencent mutuellement, soit elles sont toutes les deux soumises à un même phénomène inobservé). L'économétrie spatiale modélise cette dépendance spatiale. Plusieurs formes d'interactions existent, relatives à la variable à expliquer, aux variables explicatives ou aux variables inobservées. De nombreux modèles se retrouvent donc en concurrence, à partir d'une même définition préalable des relations de voisinage. Le chapitre 6 : *Économétrie spatiale, modèles courants* détaille la méthodologie pas à pas de choix de modèle (estimation et tests), ainsi que les précautions à prendre dans l'interprétation des résultats. La façon dont les modèles d'économétrie spatiale peuvent être appliqués à l'étude des données de panel est présentée dans le chapitre 7 : *Économétrie spatiale sur données de panel*.

L'hétérogénéité spatiale désigne le fait que l'influence des variables explicatives sur la variable dépendante varie avec la localisation des observations. La régression géographiquement pondérée ou le lissage spatial prennent en compte ce phénomène. Indépendamment d'un modèle de régression, le *lissage spatial* (chapitre 8) filtre l'information pour révéler les structures spatiales sous-jacentes. La *régression géographiquement pondérée* (chapitre 9) répond plus précisément au constat qu'un modèle de régression estimé sur l'ensemble d'un territoire d'intérêt peut ne pas appréhender de façon adéquate les variations locales. La régression géographiquement pondérée permet, notamment à l'aide de représentations cartographiques associées, de repérer où les coefficients locaux s'écartent le plus des coefficients globaux, et de construire des tests permettant d'apprécier si et comment le phénomène est non stationnaire.

Qu'elles soient destinées à prendre en compte la dépendance ou l'hétérogénéité spatiale, les méthodes d'analyse spatiale ont été développées à partir de données exhaustives. Elles peuvent cependant enrichir l'éventail des techniques liées aux sondages. Ces techniques sont particulièrement importantes pour les instituts de statistique publique, dont les données sont souvent obtenues grâce à une enquête. En amont, la constitution des entités à sélectionner aux premiers degrés d'un plan de sondage et la sélection de l'échantillon peuvent être améliorées grâce aux techniques d'échantillonnage spatial présentées dans le chapitre 10. En aval, le chapitre 11 : *Économétrie spatiale sur données d'enquête* présente les écueils liés à l'estimation d'un modèle d'économétrie spatiale sur données échantillonnées et évalue les potentielles corrections proposées par la littérature empirique. Le chapitre 12 : *Estimation sur petits domaines et corrélation spatiale* présente les méthodes petits domaines et la façon dont la prise en compte de la corrélation spatiale peut améliorer les estimations.

La quatrième partie du manuel : *Prolongements* introduit deux chapitres qui utilisent directement la dimension spatiale des données, tout en s'éloignant du traitement classique de la dépendance ou l'hétérogénéité spatiale. L'analyse des réseaux permet de prendre en compte l'ensemble des flux entre les territoires pour déterminer les relations privilégiées. Les techniques *d'analyse et de partitionnement de graphes* sont présentées dans le chapitre 13. La profusion de données géocodées va de pair avec un risque de divulgation élevé, puisque le nombre de variables nécessaires pour identifier une personne de manière unique diminue considérablement lorsque l'individu auteur de l'intrusion connaît la position géographique précise. Ce sujet est crucial pour les instituts de statistiques qui sont soumis à de fortes demandes de diffusion de données sensibles à des niveaux géographiques toujours plus fins. Le chapitre 14 : *Confidentialité des données spatiales* vise à proposer des suggestions pour évaluer et gérer le risque de divulgation, tout en préservant les corrélations spatiales.

La lecture des trois premiers chapitres est recommandée pour faciliter la compréhension de l'ensemble du manuel. Le préambule de chaque chapitre précise ensuite les chapitres particuliers dont la lecture préalable est nécessaire à la bonne compréhension du chapitre. Le corps du texte

présente la théorie fondamentale et les exemples d'application pratique. Les encadrés sont des extensions plus techniques dont la lecture n'est pas impérative pour comprendre l'essentiel de la méthode.

Références - Editorial Insee

ANSELIN, Luc (2013). *Spatial econometrics : methods and models*. T. 4. Springer Science & Business Media.

BIVAND, Roger S., Edzer PEBESMA et Virgilio GOMEZ-RUBIO (2008). *Applied spatial data analysis with R*. Springer.

BRUNSDON, Chris et Lex COMBER (2015). *An Introduction to R for Spatial Analysis Et Mapping*. Sage London.

CALOZ, Régis et Claude COLLET (2011). *Analyse spatiale de l'information géographique*. PPUR Presses polytechniques.

CRESSIE, Noel (1993a). *Statistics for spatial data*. John Wiley & Sons.

FISCHER, Manfred M et Arthur GETIS (2009). *Handbook of applied spatial analysis : software tools, methods and applications*. Springer Science & Business Media.

FLOCH, Jean-Michel (2013). « Détection des disparités socio-économiques, l'apport de la statistique spatiale ».

GELFAND, Alan E et al. (2010). *Handbook of spatial statistics*. CRC press.

HAINING, Robert P (2003). *Spatial data analysis : theory and practice*. Cambridge University Press.

KRUGMAN, Paul R (1991). *Geography and trade*. MIT press.

PACE, R Kelley et JP LESAGE (2009). « Introduction to spatial econometrics ». *Boca Raton, FL : Chapman &Hall/CRC*.

SCHABENBERGER, Oliver et Carol A GOTWAY (2004). *Statistical methods for spatial data analysis*. CRC press.

ZANINETTI, Jean-Marc (2005). *Statistique spatiale : méthodes et applications géomatiques*. Hermès science publications.