*S*éminaire de *M*éthodologie *S*tatistique

Jeudi 17 mai 2018, 14h15 - 17h15,
Insee - Bâtiment White
88 avenue Verdier - 92120 Montrouge
Salle - Malinvaud-Closon (RC-B-224 et RC-B-228)

Présentation du manuel de statistique spatiale

Afin de tenir compte des évolutions récentes dans l'accès aux données géographiques et de la demande croissante d'information statistique finement localisée, l'Insee a entrepris une refonte de son système d'information géographique, notamment pour réaliser et exploiter le recensement de la population. Pour ce faire, l'Insee s'appuie sur les nombreuses initiatives internationales destinées à mieux coordonner les informations statistiques et géographiques.

Dans ce contexte, la division des méthodes et référentiels géographiques a proposé à Eurostat et au Forum Européen de Statistique et de Géographie (EFGS) la rédaction collaborative d'un manuel de statistique spatiale. Le manuel s'inscrit dans une volonté générale, soutenue par Insee 2025, d'améliorer la résolution spatiale des statistiques produites par le SSP. Il décrit un ensemble de méthodes statistiques mobilisables à partir de données géolocalisées. De la description des données spatiales jusqu'à la gestion de leur confidentialité en passant par l'économétrie spatiale, l'échantillonnage ou le lissage spatial, le manuel couvre un large spectre de sujets. A l'occasion de la parution du manuel, le séminaire sera l'occasion d'en présenter plusieurs chapitres, illustrant la variété des thématiques abordées.

- **Introduction et présentation du manuel**
  - **Marie-Pierre de Bellefon, Vincent Loonis (Insee)**
- **Codifier la structure des données et indices d'auto corrélation**
  - **Marie-Pierre De Bellefon, Vincent Loonis, Ronan Le Gleut (Insee), Salima Bouayad Agha (GAINS (TEPP) et CREST Le Mans Université)**
- **Les configurations de points**
  - **Jean-Michel Floch (Insee), Eric Marcon (AgroParisTech, UMR EcoFoG), Florence Puech (RITM, Univ. Paris-Sud, Université Paris-Saclay & CREST)**
- **Prendre en compte l'hétérogénéité spatiale avec le lissage spatial**
  - **Laure Genebes, Auriane Renaud et François Sémécurbe (Insee)**
- **Prendre en compte la dépendance spatiale avec l'économétrie spatiale**
  - **Salima Bouayad Agha (GAINS (TEPP) et CREST Le Mans Université), Julie Le Gallo , Lionel Védrine (CESAER, AgroSup Dijon, INRA, Université de Bourgogne Franche-Comté)**
- **Confidentialité des données spatiales : comment diffuser des données protégées sans invalider les analyses locales ?**
  - **Maël-Luc Buron, Maëlle Fontaine (Insee)**

Résumé des interventions

Codifier la structure des données et indices d'auto corrélation

Marie-Pierre De Bellefon, Vincent Loonis, Ronan Le Gleut (Insee), Salima Bouayad Agha (GAINS (TEPP) et CREST Le Mans Université)

La localisation relative des données joue un rôle dans la compréhension et l'analyse de nombreux phénomènes et doit donc être prise en compte. Les domaines d'application de la statistique spatiale sont nombreux et sont autant de domaines pour lesquels les données ont en commun d'être localisées dans l'espace géographique et d'être ni indépendantes ni identiquement distribuées. Avant de pouvoir mesurer l'intensité des relations spatiales entre agents ou entre zones géographiques, et de mettre en œuvre les méthodes statistiques les plus adaptées il faut définir *a priori* leurs relations de voisinage. Celles-ci doivent être au plus près des véritables interactions spatiales entre les entités considérées. A la différence de données temporelles pour lesquelles les relations sont séquentielles, les données spatiales se caractérisent par des relations multidirectionnelles et multilatérales. La codification des relations de voisinage repose sur des critères de contiguïté ou de distance entre les observations. Nous présenterons dans un premier temps la démarche globale de codification des relations de voisinage spatiales illustrée par quelques exemples concrets avant d'aborder la notion de poids qu'il faut accorder aux voisins de manière à exprimer la dépendance spatiale. Dans un second temps, nous présenterons les indices d'auto corrélation spatiale qui vont permettre de mesurer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace, étape indispensable avant d'envisager toute spécification des interactions spatiales dans un modèle approprié. Ces indices peuvent être mesurés au niveau global et/ou local.

Les configurations de points

Jean-Michel Floch (Insee), Eric Marcon (AgroParisTech, UMR EcoFoG), Florence Puech (RITM, Univ. Paris-Sud, Université Paris-Saclay & CREST)

L'analyse des données spatialisées sur des mailles prédéfinies pose des problèmes que les géographes regroupent sur le terme de MAUP (Modifiable areal unit problem). Les données localisées étant de plus en plus abondantes (implantation sectorielle d'établissements industriels ou commerciaux, localisation des établissements scolaires au sein des villes...), il est tentant de vouloir préserver la richesse des données individuelles et de travailler en conservant la position exacte des entités étudiées. Si tel est le cas, des analyses statistiques peuvent être élaborées à partir de données géolocalisées sans procéder à une quelconque agrégation géographique. Les observations sont alors appréhendées comme des points dans l'espace et l'objectif est de caractériser ces distributions de points. Sont-elles distribuées « au hasard » (un hasard qu'il faut définir), de façon régulière ou manifestent-elles des tendances à l'agrégation ? Dans notre intervention, nous explicitons les méthodes statistiques pouvant être mobilisées pour traiter des données ponctuelles spatialisées : définition des principaux concepts et présentation des indicateurs les plus pertinents. Différents exemples sur données réelles sont présentés pour illustrer notre propos. Pouvoir traiter de telles informations individuelles spatialisées permet de travailler sur des données qui sont aujourd'hui de plus en plus accessibles et recherchées car elles permettent des analyses très précises sur les comportements des acteurs économiques.

Prendre en compte l'hétérogénéité spatiale avec le lissage spatial
Laure Genebes, Auriane Renaud et François Sémécurbe (Insee)

L'objectif de cette intervention est de présenter de façon pédagogique les méthodes de lissage à noyau. Dans un premier temps, nous présenterons les limites des cartes choroplèthes (carte de densité). Nous nous attarderons, en particulier, sur la dépendance des résultats au choix de la maille géographique de restitution, autrement dit au 'Modifiable Areal Unit Problem'. Ensuite, nous introduirons progressivement la notion de noyau qui vient en quelque sorte se substituer aux unités géographiques des cartes choroplèthes. Puis, nous montrerons que la logique du lissage peut-être généralisée à d'autres types de cartes et d'indicateurs. Enfin, nous terminerons notre présentation en évoquant la question du choix du rayon de lissage. L'estimation de l'intensité d'un processus ponctuel sera évoquée en tant que référence théorique, mais ne sera pas développée dans cette présentation.

Prendre en compte la dépendance spatiale avec l'économétrie spatiale
Salima Bouayad Agha (GAINS (TEPP) et CREST, Le Mans Université), Julie Le Gallo, Lionel Védrine (CESAER, AgroSup Dijon, INRA, Université de Bourgogne Franche-Comté)

Nous proposons une présentation synthétique des méthodes d'économétrie spatiale appliquées aux données de panel. Nous insistons principalement sur les spécifications et les méthodes implémentées dans le package "splm" disponible sous R. Une analyse de la deuxième loi de Verdoorn illustrera cette présentation. Enfin, nous présenterons des extensions récentes des modèles spatiaux sur données de panel.

Confidentialité des données spatiales : comment diffuser des données protégées sans invalider les analyses locales ?
Maël-Luc Buron, Maëlle Fontaine (Insee)

Les méthodes de traitement de la confidentialité font aujourd'hui l'objet d'une littérature abondante et d'outils bien rodés. Cependant, la plupart de ces méthodes ne prennent pas en compte explicitement l'information géographique désormais contenue dans les fichiers à traiter. Or, le risque de cet oubli est de déformer fortement les corrélations spatiales dans les données diffusées, conduisant ainsi à une forte baisse d'utilité pour les utilisateurs souhaitant effectuer des analyses à un niveau fin. Cet enjeu est d'autant plus important à l'heure où les instituts nationaux de statistique sont amenés à diffuser de plus en plus de données infra-communales, en particulier carroyées. Dans notre présentation, une première partie se focalisera sur la notion de risque de divulgation : comment le définit-on et le mesure-t-on, et pourquoi celui-ci augmente en présence de données spatiales ? Une deuxième partie présentera de façon succincte différentes méthodes de traitement de la confidentialité prenant en compte la géographie. Enfin, une troisième et dernière partie présentera un exemple applicatif concret autour d'une diffusion de données ménages sous forme carroyée. Dans cette partie, nous pourrons évoquer les résultats des prolongements de travaux qui ont eu lieu depuis la rédaction du chapitre.