Introduction aux méthodes quantitatives en sciences sociales avec R

Philippe Apparicio et Jérémy Gelb

2020-11-11

Table des matières

Liste des tableaux

Table des figures

Préface

Comment lire ce livre

Si vous googlez l'expression « comment lire un livre ? », vous trouverez une multitude de conseils et astuces. Pour ce livre, nous conseillons de le lire de gauche à droite et page par page. Plus sérieusement, il comprend plusieurs types de blocs de texte qui, on l'espère, faciliteront la lecture.

Bloc packages : habituellement localisé en début du chapitre, il comprend la liste des packages R utilisés pour un chapitre.

Bloc objectif : comprend une description des objectifs d'une section.

Bloc notes : comprend une information secondaire sur une notion, un élément, une idée abordée dans une section.

Bloc pour aller plus loin : peut comprendre des références ou des extensions d'une méthode statistique abordée dans une section.

Bloc astuce : décrit un élément qui vous facilera le vie : une propriété statistique, un *package*, une fonction, une syntaxe R.

Bloc attention : comprend une notion ou un élément important à bien maîtriser.

Structure du livre

À écrire plus tard.

Remerciements

Note au beau Cargo (chien Mira) qui nous supporte dans l'écriture du livre !

Fig. 1 : Cargo, le plus beau

À propos des auteurs

Philippe Apparicio (http://www.ucs.inrs.ca/philippe-apparicio) est professeur titulaire au Centre Urbanisation Culture Société de l'INRS (http://www.ucs.inrs.ca/). Il enseigne au programme de maîtrise en études urbaines (http://www.ucs.inrs.ca/ucs/etudier/programmes/etudes-urbaines) les cours *méthodes quantitatives appliquées aux études urbaines* et *analyses spatiales appliquées aux études urbaines*. Il a aussi créé et enseigné, il y a plusieurs années, le cours *systèmes d'information géographique appliqués aux études urbaines*. Durant les dernières années, il a offert plusieurs formations aux Écoles d'été du Centre interuniversitaire québécois de statistiques sociales (CIQSS, https://www.ciqss.org/). Titulaire de la Chaire de recherche du Canada (niveau 2) sur l'équité environnementale et la ville, il est le directeur du **laboratoire d'équité environnementale** (http://laeq.ucs.inrs.ca). Géographe de formation, ses intérêts de recherche actuels incluent la justice et l'équité environnementale, la pollution atmosphérique et le bruit et le vélo en ville. Il a publié une centaine d'articles dans différents domaines des études urbaines et de la géographie.

Jérémy Gelb est candidat au doctorat en études urbaines à l'INRS (sous la supervision de Philippe Apparicio) et membre du **laboratoire d'équité environnementale** (http://laeq.ucs.inrs.ca). Son sujet de thèse porte sur l'exposition des cyclistes aux pollutions atmosphériques et sonores en milieu urbain. Il utilise quotidiennement des systèmes d'information géographique (SIG) et est tombé dans la marmite de l'*open source* avec le triptyque QGIS, R et Python au début de sa maîtrise. Il a récemment développé deux packages R : **geocmeans** et **spNetwork**, permettant respectivement d'effectuer des analyses de classification floue non-supervisée pondérée spatialement et des estimations de densité par kernel sur réseau.

Philippe et Jérémy travaillent étroitement ensemble depuis déjà plusieurs années. Avec d'autres collègues, ils ont copubliés plusieurs articles (**??????????**). Tous deux s'intéressent à l'exposition des cyclistes à la pollution atmosphérique et sonore dans plusieurs villes à travers le monde : Philippe ayant une préférence pour les collectes dans les villes du Sud Global (notamment indiennes, africaines et latino-américaines) et Jérémy dans les villes du Nord (européennes et nord-américaines).

Chapitre 1

Régressions régression multiniveaux

Dans le précédent chapitre, nous avons abordé les modèles à effets mixtes qui permettent d'introduire à la fois des effets fixes et des effets aléatoires (GLMM). Dans ce chapitre, nous poursuivons sur cette voie avec une nouvelle extension des modèles GLM : les modèles multiniveaux. Ces modèles sont simplement une extension des modèles à effets mixtes et permettent de modéliser un phénomène avec une structure hiérarchique des données, tel que décrit dans le chapitre précédent.

Rappel de la structure hiérarchique des données

Exemple à deux niveaux : il s'agit de modéliser un phénomène y_{ij}, soit une variable dépendante Y pour un individu i (niveau) niché dans un groupe j (niveau 2). Par exemple, modéliser l'indice de masse corporelle (IMC) de 5000 individus résidant dans 100 quartiers différents.

Exemple à trois niveaux : il s'agit de modéliser un phénomène y_{ijk}, soit une variable dépendante Y pour un individu i (niveau 1), niché dans un groupe j (niveau 2) appartenant à un groupe k (niveau 3). Par exemple, modéliser les notes à un examen de mathématiques d'élèves (niveau 1) nichés dans des classes (niveau 2) nichées dans des écoles (niveau 3).

Nous avons largement décrits précédemment trois principaux types de modèles d'effets mixtes (GLMM) :

- les GLMM avec constantes aléatoires qui permettent d'avoir une constante différente pour chacun des groupes (niveau 2).
- les GLMM avec pentes aléatoires qui permettent de faire varier une variable indépendante au niveau 1 (coefficient) en fonction des groupes au niveau 2.
- les GLMM avec constantes et pentes aléatoires.

Les modèles multiniveau se différencient des modèles à effets mixtes puisqu'ils permettent d'introduire des variables indépendantes mesurées aux niveaux supérieurs (2 et 3).

1.1 Les modèles multiniveaux : deux intérêts majeurs

Les modèles multiniveaux ont deux principaux avantages : la répartition de la variance entre les différents niveaux et l'introduction de variables explicatives aux différents niveaux du modèle.

1.1.1 La répartition de la variance entre les différents niveaux

Les modèles multiniveaux permettent d'estimer comment se répartit la variance entre les différents niveaux du jeu de données. Dans les deux exemples de l'encadré précédent, ils permettraient de répondre

aux questions suivantes :

- Quel niveau explique le plus l'IMC, le niveau individuel (niveau 1) ou le niveau contextuel (niveau 2) ?
- Comment se répartit la variance des notes à l'examen de mathématiques entre les trois niveaux ? A-t-on plus de variance pour les individus (niveau 1) ou au sein des classes (niveau 2) ou entre les différentes écoles (niveau 3) ?

1.1.2 L'estimation des coefficients aux différents niveaux

Les modèles multiniveaux permettent d'estimer simultanément les coefficients de plusieurs variables indépendantes introduites à chacun des niveaux du modèle. Autrement dit, de voir comment les variables indépendantes introduites aux différents niveaux influencent la variable Y dépendante mesurée au niveau 1. Si nous reprenons l'exemple à trois niveaux (élèves / classes / école), plusieurs facteurs peuvent influencer la réussite ou la performance scolaire des élèves aux différents niveaux :

- **Variables indépendantes au niveau 1** (élève) : âge, sexe, statut socioéconomique, langue maternelle autre que la langue d'enseignement…
- **Variables indépendantes au niveau 2** (classe) : nombre d'élèves par classe, programme spécialisée ou pas…
- **Variables indépendantes au niveau 3** (école) : indice de défavorisation de l'école, école publique ou privée, qualité des infrastructures de l'école (bâtiment, gymnase, cour d'école)…

Dans la même veine, afin d'illustrer l'apport des modèles multiniveau dans le champ de la géographie de la santé, Philibert et Apparicio (**?**, p. 129) signalent que « pour un modèle à deux niveaux, il s'agit de modéliser y_{ij}, par exemple l'IMC d'un individu i (niveau 1) résidant dans un quartier j (niveau 2). Il est alors possible de mettre des variables explicatives tant au niveau 1 (âge, sexe, revenu, niveau d'éducation, etc.) qu'au niveau 2 (niveau de défavorisation sociale du quartier, offre de services et d'équipements sportifs et récréatifs, caractéristiques de l'environnement urbain, etc.). Dans cet exemple, on peut voir comment la modélisation multiniveaux permet d'estimer simultanément les effets environnementaux et individuels de manière à distinguer la contribution de chacun des niveaux (ex. : l'effet du revenu des individus et celui de la défavorisation du quartier) dans l'explication des variations géographiques observées ».

Évaluer les effets de milieu avec des analyses multiniveaux

Dans les champs de la santé des populations et des études urbaines, les modèles multiniveaux sont largement mobilisés pour évaluer les effets de milieu (*neighbourhoods effects* ou *area effects* en anglais).

Atkinson et Kintrea (**?**, p. 2278) définissent les « effets de milieu comme le changement net dans les potentialités de l'existence (*life chances*) attribuable au fait de vivre dans un quartier (ou une zone) plutôt qu'un autre » (traduction libre). Les effets de milieu peuvent être positifs ou négatifs et peuvent concernant aussi bien les enfants que les adultes.

Les analyses multiniveaux sont particulièrement adaptées à l'évaluation des effets de milieu. En effet, plusieurs phénomènes – état de santé, comportement ou choix individuels – peuvent être en effet influencés à la fois par des caractéristiques individuelles (âge, sexe, niveau de revenu, niveau d'éducation, etc.) et par des caractéristiques contextuelles (caractéristiques du quartier).

Avec un modèle multiniveau, une fois contrôlées les caractéristiques individuelles (variables indépendantes mesurées au niveau 1), il est alors possible d'évaluer l'effet des caractéristiques du quartier (variables indépendantes mesurées au niveau 2) sur un phénomène y_{ij} mesuré pour un individu i résidant dans un quartier j.