Génération de données synthétiques spatio-temporelles pour l'épidémiologie environnementale

Amadou-Tidiane Niang[1,*] Aghiles Hamroun[2,3] Camille Ternynck[1,2] Michaël Génin[1,2]

[1]Univ Lille, ULR 2694 – METRICS : Évaluation des technologies de santé et des pratiques médicales, F-59000 Lille, France
[2]CHU de Lille, [3]UMR1167 RID-AGE
[*]E-mail : amadou-tidiane.niang@univ-lille.fr

Introduction. Le partage de données de cartographie des maladies spatio-temporelles à l'échelle de petites zones devient de plus en plus difficile : au-delà des contraintes éthiques et réglementaires, les effectifs faibles et les événements rares peuvent engendrer des risques importants de ré-identification lors de la diffusion de cartes ou de tableaux. Cette tension est particulièrement forte en épidémiologie environnementale, où les questions scientifiques reposent sur une résolution spatiale et temporelle fine afin de détecter des variations localisées, d'évaluer les évolutions temporelles et d'éclairer la décision en santé publique. Dans ce contexte, les données synthétiques constituent un compromis attractif, mais seulement si elles conservent une grande utilité pour les analyses ultérieures. En pratique, cette utilité exige plus que la reproduction des tendances nationales : les jeux de données synthétiques doivent préserver la structure spatio-temporelle locale (y compris les agrégats spatiaux et la dynamique temporelle) et rester compatibles avec l'inférence écologique, en maintenant les associations avec les principaux déterminants territoriaux tels que les expositions environnementales et le contexte socio-économique.

Méthodes. Nous proposons une chaîne de production de données synthétiques spatio-temporelles entièrement bayésienne, fondée sur des modèles standards de cartographie des maladies ajustés avec INLA. Les effectifs observés sont modélisés par des spécifications hiérarchiques de Poisson avec un décalage (offset) pour les effectifs attendus, et une structure latente combinant une composante spatiale BYM2, une tendance temporelle lisse et un terme d'interaction espace-temps. En amont de la synthèse, nous sélectionnons la structure de dépendance spatio-temporelle parmi plusieurs spécifications concurrentes à l'aide de critères usuels de comparaison de modèles bayésiens, puis nous intégrons des covariables écologiques (par exemple la pollution de l'air, des indicateurs d'occupation des sols, la défavorisation, l'accessibilité aux soins) dans le modèle retenu. Les effectifs synthétiques sont générés par un synthétiseur multinomial prédictif a posteriori : nous effectuons de multiples tirages dans la distribution a posteriori du modèle ajusté, traduisons chaque tirage en une surface de risque spatio-temporelle, puis simulons les effectifs conditionnellement aux totaux observés. Cela produit un ensemble de jeux de données synthétiques qui propage naturellement l'incertitude du modèle bayésien. L'utilité est évaluée par des diagnostics calculés sur l'ensemble des réplicats, portant sur la préservation des effectifs bruts, de la structure spatio-temporelle du risque, des profils de probabilité de dépassement et des signaux de régression écologique. Pour la diffusion, un jeu de données unique peut être choisi de manière pragmatique en minimisant une fonction objectif qui cible un profil diagnostique « idéal ».

Résultats. Notre approche a été appliquée aux données d'incidence spatio-temporelle de la maladie rénale issues du registre national français REIN (2012-2021), en intégrant des déterminants écologiques incluant la prévalence du diabète, l'accès aux soins, la défavorisation à l'échelle des zones et la pollution de l'air ambiant ($PM_{2,5}$). Le synthétiseur prédictif a posteriori a produit des jeux de données synthétiques reproduisant fidèlement les totaux annuels, préservant les principales structures spatiales et spatio-temporelles, et maintenant les associations clés avec

les covariables écologiques intégrées au modèle. L'évaluation des diagnostics sur les réplicats prédictifs a posteriori a permis de quantifier de façon transparente la fidélité ainsi que la variabilité induite par la synthèse, et a appuyé la sélection d'un jeu de données unique représentatif en vue de son partage.

Conclusion. Nous présentons un cadre bayésien spatio-temporel pour générer des données synthétiques de maladies à l'échelle de petites zones, qui préserve les structures espace-temps et les relations d'exposition écologique au cœur de l'épidémiologie environnementale. En combinant la synthèse prédictive a posteriori, des diagnostics fondés sur les distributions et une étape de sélection basée sur une fonction objectif, la méthode offre une voie pratique pour diffuser un jeu de données synthétique unique tout en conservant une forte utilité scientifique.

Mots-clés : Cartographie spatio-temporelle des maladies ; Données de santé synthétiques à petite échelle ; Épidémiologie environnementale