De l'apport des techniques spatiales à la compréhension des liens entre environnements et maladies

Gilles MAIGNANT I maignant@unice.fr

CNRS, UMR 1252 SESSTIM, Marseille

Mots clés : **Approche spatiale, cartogramme, environnement, exposition, maladies**

Comprendre la distribution spatiale des maladies, notamment les infarctus du myocarde ou les leucémies n'est pas si aisée. En effet, de très nombreux facteurs de voisinages, qu'ils soient sociodémographiques, environnementaux, d'accès aux transports etc., peuvent jouer un rôle majeur dans cette répartition. Au-delà des facteurs eux-mêmes, cette note a pour objectif de s'intéresser à deux techniques différentes permettant d'aborder cette question. La première s'intéresse à une approche guidée par les données pour comprendre l'influence du voisinage sur la distribution spatiale des infarctus du myocarde (cf. Kihal-Talantikite et al.) tandis que la deuxième questionne la significativité statistique* des clusters de maladies à l'aide de cartogrammes (cf. Kronenfeld et al.). Ces deux articles s'insèrent plus généralement dans le champ de la géographie de la santé.

Développement de méthodes spatiales s'appuyant sur des données pour évaluer l'influence du voisinage sur la distribution spatiale des infarctus du myocarde.

KIHAL-TALANTIKITE W. et al. (2017). Developing a data-driven spatial approach to assessment of neighbourhood influences on the spatial distribution of myocardial infarction. International Journal of Health Geographics, 16 : p.22.

Résumé

De plus en plus d'études montrent que le voisinage (via différents facteurs) influence notre état de santé. Des caractéristiques variées tels que l'environnement socio-économique, l'accessibilité à des aménités (centres de loisirs, bibliothèques, parcs et transports), la cohésion sociale peuvent, une fois combinées, contribuer à augmenter les inégalités de santé. L'objectif de cet article est de coupler des méthodes spatiales orientées données avec des techniques d'analyses de clusters, pour décrire la distribution spatiale du risque d'infarctus du myocarde et les caractéristiques contextuelles associées à cette distribution, dans l'aire métropolitaine de Strasbourg (AMS), représentant une superficie de 316 km^2, 33 municipalités et 190 IRIS*. Tous les événements d'infarctus du myocarde pour la période 2000-2007, de patients âgés de 35 à 74 ans ont été recensés et géocodés au bloc correspondant à l'adresse effective de recensement (anonymisation des données). D'un point de vue méthodologique, les données récoltées n'étaient pas disponibles à la même échelle spatiale, c'est pourquoi les auteurs ont construit un maillage spécifique (bloc) permettant de prendre en compte l'hétérogénéité des données et ceci pour trois raisons (stabilité des unités géographiques, prise en compte des contraintes de construction des unités spatiales administratives, homogénéisation des données contextuelles), maillage basé sur la méthode classique du plus proche voisin, avec une taille de 250 m * 250 m (maille carrée). Un algorithme de découpage des différentes zones a été utilisé pour désagréger les données à cette échelle. Les auteurs ont considéré 27 variables regroupées en trois grands domaines contextuels (socio-économiques, accessibilité à des services, environnement psychosocial). Dans un premier temps, l'approche a consisté à estimer les distances entre les unités en attribuant un poids spécifique à chaque variable (hiérarchisation des valeurs propres (ACP*, AFC*)). L'analyse spatiale a permis d'une part, d'identifier la localisation des clusters et d'autre part d'évaluer et comprendre les rôles du voisinage dans la distribution spatiale du risque d'infarctus du myocarde. L'étude a montré que l'incidence de l'infarctus du myocarde n'est pas distribuée aléatoirement d'un point de vue spatial pour notamment deux clusters bien marqués, l'un à forte incidence au Nord de l'AMS (RR*=1.70) et un à très faible incidence, situé dans la première et deuxième périphérie de Strasbourg (RR=0.04). Ces résultats montrent que le lieu où l'incidence du risque est la plus forte est caractérisé par un contexte socioéconomique défavorable malgré le fait que les habitants aient un bon accès aux loisirs et activités récréatives.

Commentaire

Cette étude dispose d'un fort apport méthodologique qui permet de coupler des données hétérogènes à une échelle très fine et d'atténuer l'effet de MAUP* tout en préservant l'anonymisation obligatoire des données de santé. Le fait de ne pas attribuer de poids *a priori* aux variables contextuelles de voisinage permet d'avoir une véritable approche guidée par les données. Mais comme pour toute étude géographique ou sociologique, les résultats sont données-dépendant, c'est-à-dire que le nombre de variables et de domaines contextuels utilisés reste limité. Par exemple, les données de bruit n'ont pu être intégrées pour des questions de sensibilité politique, bien qu'existantes. De même, les données de santé ont été collectées sur la période 2000-2008 alors que la plupart des données contextuelles ne sont pas disponibles annuellement

mais pour une année particulière (2007, 2008 ou 1999 pour les données socioéconomiques). Ce gap temporel peut engendrer une variabilité dans les résultats qui en résultent. Toutefois, l'utilisation de jeux de données, associée à une approche orientée données couplée à un SIG*, permet aux décideurs politiques d'identifier et de lutter contre les inégalités spatiales de santé, en leur fournissant une vision des disparités spatiales elles mêmes. Si on peut affirmer aux décideurs que certaines de ces disparités spatiales ont un lien de nature causale avec un phénomène de santé, tel que la survenue de l'infarctus du myocarde, et lui permettre d'identifier des cibles d'actions dont on peut penser qu'elles conduiraient à une réduction des phénomènes de santé indésirables, celles-ci feraient l'objet de politiques publiques de modification positive.

Visualisation de la significativité statistique des clusters de maladies à l'aide de cartogrammes

KRONENFELD BJ. et. al. (2017). Visualizing statistical significance of disease clusters using cartograms. International Journal of Health Geographics, vol. 16 : p.19.

Résumé

En épidémiologie, une longue tradition consiste à utiliser des cartes pour détecter des clusters de maladies. Mais celles-ci exagèrent souvent l'information visuelle donnée pour des zones à faible densité de population et à surface importante, ce qui peut masquer, à l'inverse, des potentiels élevés en zone urbaine (forte densité de population, faible surface). Pour y remédier les chercheurs préfèrent utiliser des cartogrammes* s'appuyant sur des cartes de densités de population homogènes. Mais la question de l'incertitude statistique reste présente. Pour y répondre les auteurs ont mis en place une méthode qui permet de déterminer visuellement la significativité statistique des clusters regroupant un ou plusieurs districts sur le cartogramme. Ils ont développé une formule qui permet d'estimer, pour un taux donné, la superficie minimale requise pour avoir une significativité statistique (spatiale) des régions choisies *a priori* et *a posteriori* selon certaines hypothèses de tests. L'enjeu est de permettre à l'utilisateur des cartes, de faire la distinction entre des clusters statistiquement significatifs et des zones qui présentent des forts taux de maladies mais ne sont pas statistiquement significatifs à cause de leur petite population. L'approche des auteurs est basée sur le fait que la variance est normalement une fonction inverse de l'effectif de la population et donc la significativité statistique est déterminée à partir de cet effectif et du taux de maladie observée sur chacun des districts. Pour appliquer leurs méthodes les auteurs ont choisi d'analyser l'incidence de la leucémie en Californie et montrent qu'il est possible de distinguer visuellement les régions statistiquement significatives des régions statistiquement non-significatives.

Commentaire

Les auteurs donnent tout d'abord une bibliographie très complète et commentent très clairement des méthodes et techniques de cartographie utilisables en géographie de la santé. Ils utilisent des statistiques d'analyse qui fournissent un test strict d'existence et de localisation de maxima ou minima locaux, tenant compte de la probabilité qu'un cluster donné se produise n'importe où sur la carte sous l'hypothèse nulle d'une distribution aléatoire du nombre total de cas observé sur la période étudiée. Une mesure de la compacité est également utilisée pour atténuer les problèmes introduits, lorsque des clusters à forme irrégulière sont considérés (effet de taille et significativité). Deux méthodes sont proposées dans cet article, et les avantages et les limites de chacune sont discutés. Pour la première méthode, le nombre de cas de maladie par unité géographique est utilisé comme variable pour le cartogramme, la représentation cartographique est alors proportionnelle au nombre de cas, créant ainsi des cartes en anamorphoses souvent difficiles à lire, avec le biais que cette déformation peut provenir uniquement de zones à forte densité de population ; dans la seconde méthode, c'est la population à risque qui est la variable considérée pour produire le cartogramme, ce qui réduit alors ces surreprésentations spatiales. Les auteurs ont évalué deux scénarios (zones identifiées *a priori* et *a posteriori)*. Dans le premier scénario, le nombre d'occurrences de cas de la maladie « attendu » dans la zone désignée (*a priori*) peut être déterminé à partir de sa population sous l'hypothèse nulle d'une distribution aléatoire des occurrences dans toutes les zones. Cette zone est comparée à son voisinage plutôt qu'à la carte entière afin de réduire le gap entre les régions. Dans le second scénario, les auteurs proposent une méthode permettant d'identifier les zones où les taux d'événements ou leurs significativités sont maximum (statistique de scan permettant de détecter le maximum sur une fenêtre glissante). Cette détection nécessite d'avoir recours à des outils de type simulations de Monte Carlo (nombre de fenêtres glissantes fixé au préalable). L'une des limites majeures de cette méthode est que la distribution observée est dépendante de la taille et de la forme des fenêtres glissantes, mais aussi de l'agencement spatial des centroïdes considérés. Bien qu'utilisant des valeurs statistiquement significatives pour montrer l'incertitude sur les cartes épidémiologiques, comme le font de nombreux chercheurs, il faut garder à l'esprit que, par nature, une étude exploratoire ne peut pas conduire à des conclusions aussi solides qu'une étude conçue pour être capable de valider ou infirmer un résultat préalablement obtenu dans un cadre exploratoire.

Conclusion générale

Malgré toutes les avancées scientifiques permettant de les caractériser, les inégalités de santé persistent et constituent un véritable problème de santé publique. Depuis les années 2000, les experts s'intéressent à la question du rôle de l'environnement contextuel sur la survenue de maladies. A travers deux articles, tous les deux innovants en matière d'approche spatiale et de visualisation de données de santé, les auteurs ont montré que les données de santé cartographiées peuvent être facilement mal interprétées malgré notre familiarité avec l'usage des cartes, compte tenu du fort pouvoir de suggestion d'information qu'elles contiennent. Ceci soulève des questions méthodologiques importantes (taux d'incidence, significativité statistique, représentativité, techniques de cartographie et colorimétrie, …), qui ont un impact majeur sur la qualité de la conception des cartes à fournir à la sphère décisionnelle.

GENERAL CONCLUSION

Despite all the scientific advances that make it possible to characterize them, health inequalities persist and constitute a real public health problem. Since the 2000s, experts have been addressing the issue of the role of contextual environment on the occurrence of diseases. Through these two articles, both innovative in terms of spatial approach and visualization of health data, the authors have shown that mapped health data can be easily misinterpreted despite our familiarity with maps, given the strong power of suggestion of the information it contains. This raises important methodological issues (incidence rate, statistical significativity, representativeness, mapping and colorimetric techniques ...), which have a major impact on the quality of the conception of the maps to be provided to the decision-making sphere.

Lexique

ACP : Analyse en composantes principales (variables quantitatives).

AFC : Analyse factorielle des correspondances (variables qualitatives).

Cartogramme : représentation schématique d'informations statistiques.

IRIS : Unité statistique, ilots regroupés pour l'information statistique, plus petite unité administrative française pour laquelle les données socioéconomiques et démographiques sont disponibles.

MAUP : Modifiable Area Unit Problem, proposé par Openshaw et Taylor en 1979 pour désigner l'influence du découpage spatial (effets d'échelle et effets de zonage) sur les résultats de traitements statistiques ou de modélisation.

RR : risque relatif : mesure statistique souvent utilisée en épidémiologie, estime le risque de survenue d'un événement dans un groupe par rapport à un autre.

Significativité statistique : mesure estimée du degré pour lequel un résultat est "vrai" (au sens de "représentatif de la population").

SIG : Système d'Information Géographique, logiciel permettant de cartographier et de croiser des données à références spatiales.

Publications de référence

1 **Strak MJ.** et al. (2017). Associations between lifestyle and air pollution exposure: Potential for cofounding in large administrative data cohorts. Environmental Research, 156 : p.364-73.

2 **Folino F.** et al. (2017). Associations between air pollution and ventricular arrythmias in high-risk patients (ARIA Study): a multicentre longitudinal study. Lancet Planet Health, 1 : p.58-64.

Liens d'intérêts

☒ Les auteurs déclarent n'avoir aucun lien d'intérêt