UNIVERSITE DU DROIT ET DE LA SANTE - LILLE 2

FACULTE DE MEDECINE HENRI WAREMBOURG

Année : 2017

THESE POUR LE DIPLOME D'ETAT

DE DOCTEUR EN MEDECINE

Méthode générique d'amélioration de l'interopérabilité des bases de données médicales et écologiques et application sur données françaises.

Présentée et soutenue publiquement le 20 septembre 2017 à 18h
au Pôle Recherche

Par Adrien Ghenassia

JURY

Président :

Monsieur le Professeur Alain Duhamel

Assesseurs :

Madame le Professeur Florence Richard

Madame le Docteur Marie-Hélène Metzger

Monsieur le Docteur Emmanuel Chazard

Directeur de Thèse :

Monsieur Michaël Genin

Avertissement

La Faculté n'entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Table des matières

Abréviations

AIC	Akaike index criterion
ATIH	Agence technique de l'information hospitalière
AVC	Accident vasculaire cérébral
CCAM	Classification commune des actes médicaux
CIM 10	Classification internationale des maladies, 10ème révision
COSP	Change of support problem
DAS	Diagnostic associé significatif
DIC	Deviation index criterion
DP	Diagnostic principal
DREES	Direction de la recherche, des études, de l'évaluation et des statistiques
EDI	European deprivation index
EGB	Echantillon généraliste des bénéficiaires
GPS	Global Positioning system
IGN	Institut géographique nationale
INSEE	Institut national de la statistique et des études économiques
IRDES	Institut de recherche et documentation en économie de la santé
IRIS	Ilôts regroupés pour l'information statistique
MAUP	Modifiable area unit problem
MDP	Misaligned data problem
MTUP	Modifiable temporal unit problem
ORS	Observatoire Régionale de la Santé
PMSI	Programme médicalisé des systèmes d'information
PUMA	Public use microdata area
SIG	Système d'information géographique
SIR	Standardized incidence ratio
SNIIRAM	Système national d'information inter-régime de l'assurance maladie
SPARCS	Statewide Planning and Research Cooperative System

Résumé

Introduction

La disponibilité de bases de données de grandes dimensions et le développement important de la réutilisation de données et du géoréférencement ont ouvert des perspectives dans l'analyse spatiale en santé. Cependant, les études à fine échelle avec des données écologiques et médicales sont limitées par le *change of support problem* et le manque d'interopérabilité entre unités spatiales. Les méthodes de désagrégation spatiale pour résoudre ce problème introduisent des erreurs dans l'estimation spatiale. De plus, il existe un manque de cadre d'étude dans ces situations. Nous présentons ici une méthode générique d'agrégation à 2 étapes permettant de fusionner des bases de données médicales et écologiques avec un schéma standard sans utiliser de modèles de désagrégation spatiale tout en maximisant la résolution spatiale.

Méthode

Premièrement, une table de correspondance est construite après identification d'une ou plusieurs matrices de passage. Cette table relie les unités spatiales des bases de données originales avec les unités spatiales de la base de données finale.

Deuxièmement, la table de correspondance est validée par la comparaison de variables contenues dans les 2 bases de données initiales et la vérification de la validité spatiale avec un critère de continuité spatiale et un critère de résolution spatiale.

Résultats

Nous avons utilisé cette méthode pour fusionner une base de données médicales (le programme médicalisé des systèmes d'information contenant 5644 unités spatiales) avec une base de données écologiques (issue de l'Institut National de la Statistique et des Etudes Economiques contenant 36594 unités spatiales). La table de correspondance finale aboutit à 5632 unités spatiales. Elle a été validée par la comparaison du nombre de naissances dans la

base de données médicales et dans la base de données écologiques pour chaque unité spatiale finale. La médiane [intervalle inter-quartile] de la différence relative était de 2,3% [0 ; 5,7]. Le critère de continuité spatiale était faible (2,4%) et l'indice de résolution spatiale était meilleur que pour la plupart des unités administratives françaises.

Conclusions

Notre approche innovante améliore l'interopérabilité entre bases de données médicales et écologiques et facilite l'analyse spatiale à fine échelle. Nous avons montré que les modèles de désagrégation et les larges agrégations n'étaient pas nécessairement les meilleures façons de répondre au *change of support problem*.

Abstract

Background

The availability of big data in healthcare and the intensive development of data reuse and georeferencing have opened up perspectives for health spatial analysis. However, fine-scale spatial studies of ecological and medical databases are limited by the change of support problem and thus a lack of spatial unit interoperability. The use of spatial disaggregation methods to solve this problem introduces errors into the spatial estimations. Moreover, the lack of a framework limits the research reproducibility. Here, we present a generic, two-step method for merging medical and ecological databases with a standard framework that avoids the use of spatial disaggregation methods, while maximizing the spatial resolution.

Method

Firstly, a mapping table is created after one or more transition matrices have been defined. The latter link the spatial units of the original databases to the spatial units of the final database. Secondly, the mapping table is validated by (i) comparing the covariates contained in the two original databases, and (ii) checking the spatial validity with a spatial continuity criterion and a spatial resolution index.

Results

We used our novel method to merge a medical database (the French national diagnosis-related group database, containing 5,644 spatial units) with an ecological database (produced by the French National Institute of Statistics and Economic Studies, and containing with 36,594 spatial units). The mapping table yielded 5632 final spatial units. The mapping table's validity was evaluated by comparing the number of births in the medical database and the ecological databases in each final spatial unit. The median [interquartile range] relative difference was 2.3% [0; 5.7]. The spatial continuity criterion was low (2.4%), and the spatial resolution index was greater than for most French administrative areas.

Conclusions

Our innovative approach improves interoperability between medical and ecological databases and facilitates fine-scale spatial analyses. We have shown that disaggregation models and large aggregation techniques are not necessarily the best ways to tackle the change of support problem.

Introduction générale

I. Etudes écologiques

A. Définition

Une étude écologique est une étude épidémiologique dans laquelle l'évènement de santé et les facteurs de risque potentiels sont obtenus collectivement sur les mêmes populations. C'est dire que l'unité statistique n'est pas l'individu mais une population (1,2). Morgenstern liste plusieurs intérêts aux études écologiques en épidémiologie tels que le contournement des difficultés liées aux enquêtes individuelles pour la récupération de données socio-économiques ou environnementales, les difficultés liées aux problèmes d'anonymisation, la simplicité de ce type d'étude ainsi que son coût moins élevé (2). Enfin, le caractère écologique est particulièrement adapté lorsque les variations géographiques d'une variable sont supérieures aux variations individuelles (1,2). Cependant, les spécificités liées à ce type d'étude ont été discutées dans la littérature et doivent être prises en compte afin de réaliser des analyses de qualité (3,4).

B. Particularités

Les données écologiques sont des données agrégées pour lesquelles les unités statistiques sont des populations. Il est possible de transformer des données individuelles en données écologiques par une agrégation (Tableau 1). Le mode de calcul des variables agrégées est adapté au type de données : somme, médiane, taux…

Tableau 1 : Transformation de données individuelles en données écologiques par agrégation obtenu par calcul d'une somme.

Unité statistique classique	Variable individuelle	Agrégation	Unité statistique écologique	Variable écologique
Patient A	1		Groupe 1	2
Patient B	0			
Patient C	1			
Patient D	0			

Les données écologiques peuvent être tout aussi bien médicales, socio-économiques, environnementales etc. On peut citer par exemple :

- Nombre de décès par département ;
- Revenu médian par commune ;
- Indice de défaveur sociale tel que le French European Deprivation Index (French EDI) par Îlots Regroupés pour l'Information Statistique (Iris) (5);
- Type d'occupation des sols par canton ;
- Nombre de sites pollués par commune.

Morgenstern rappelle en 1982 les limites qui se posent dans les études épidémiologiques écologiques (6). La principale limite est celle du biais écologique décrit par Selvin ainsi que Robinson (7,8). Morgenstern distingue 2 composantes dans le biais écologique :

- Le biais d'agrégation venant du regroupement d'individus. Dans ce cas, les individus ayant présentés un évènement ne sont pas forcément les individus ayant été exposés.
- Le biais de spécification, comparable au biais de confusion en épidémiologie classique. Supposons par exemple une zone géographique avec une fréquence d'asthme plus élevée. La mise en évidence de cette zone peut aussi bien être une mise en évidence d'une zone de pollution atmosphérique plus importante qu'une mise en

évidence d'une zone défavorisée avec une fréquence plus élevée de logements insalubres ou encore correspondre à un regroupement de personnes présentant une fréquence plus élevée d'allergies.

Ainsi, les conclusions d'une étude écologique ne permettent jamais de tirer des conclusions à l'échelle individuelle. Cependant, la maitrise des paramètres de l'étude permet le contrôle de ces biais écologiques (9) :

- La constitution de groupe le plus homogène possible qui permet de limiter le biais d'agrégation ;
- La constitution de groupe de taille la plus faible possible limitant là aussi le biais d'agrégation ;
- L'utilisation de méthodes statistiques permettant l'ajustement sur des facteurs de confusion potentiels et ainsi la limitation du biais de spécification.

Ces études écologiques ne rentrent pas nécessairement dans le cadre de l'analyse spatiale. A l'inverse, un grand nombre d'analyses spatiales en santé repose sur les principes des études écologiques.

II. Analyse spatiale en santé

Melnick pose en 1999 le problème en ces termes : « Parmi les 3 principales variables épidémiologiques temps, localisation, personne, la localisation a toujours été la plus difficile à analyser et décrire. » (10). Pourtant la localisation est d'un intérêt majeur en santé car cette caractéristique permet à travers l'analyse spatiale d'évaluer la répartition d'un événement de santé sur le territoire et de le relier à des facteurs de risques écologiques.

Il est notamment intéressant de voir l'attention importante que nécessite la manipulation de données spatiales dont le système de coordonnées a un impact direct sur les coordonnées géographiques utilisées pour les localiser dans l'espace. Certains systèmes de coordonnées reposent sur la sphère (le système du *Global Postioning System*, GPS, dont le système de coordonnées est appelé WGS84 par exemple), d'autres sur des projections planes qui ont la propriété de conserver les angles (projections de type Mercator par exemple) ou les distances (projections de type Lambert par exemple) mais jamais les deux (11,12). De plus, ces données peuvent être des zones (une commune), des points (une station de mesure) ou des lignes (une ligne électrique). Tout d'abord un court historique permettra de revenir sur le développement de l'analyse spatiale en santé afin de comprendre comment les outils informatiques modernes ont permis le développement de ce type d'analyse puis les différents types de données manipulées serons vus en détail.

A. Développement de la discipline

La première analyse spatiale en santé connue et documentée est le travail du docteur John Snow sur la transmission de cas de choléra dans le quartier de Soho à Londres, publié en 1855 (13). Le choléra est une maladie fréquente au XIXème siècle mais dont l'origine n'est pas connue. La symptomatologie est une diarrhée profuse pouvant entrainer une déshydratation sévère. Paccini met en évidence l'existence du *Vibrio cholerae*, mais il faudra attendre 1884 pour que Robert Koch fasse formellement le lien entre *Vibrio cholerae* et la pathologie (14,15).

A l'époque de John Snow, la théorie des miasmes prédomine. Il s'inscrit en opposition à cette thèse en soutenant la transmission oro-fécale, en particulier par l'eau. Sa démarche expérimentale sera de répertorier l'ensemble des cas du quartier de Soho à Londres sur une carte selon leur lieu d'habitation. Il observe de façon empirique l'existence d'un agrégat

spatial de cas de choléra autour de la rue Broad Street. Ces patients du fait de leur localisation partagent l'exposition à un facteur de risque compatible avec la transmission oro-fécale de la maladie : une pompe à eau publique. A l'issue de ses recherches, la fermeture de la pompe sera proposée et l'incidence du choléra dans le quartier de Soho va rapidement diminuer. John Snow utilise l'analyse spatiale pour cartographier une pathologie afin d'observer sa distribution et ainsi la relier à un facteur de risque environnemental dans l'objectif de conforter son hypothèse initiale.

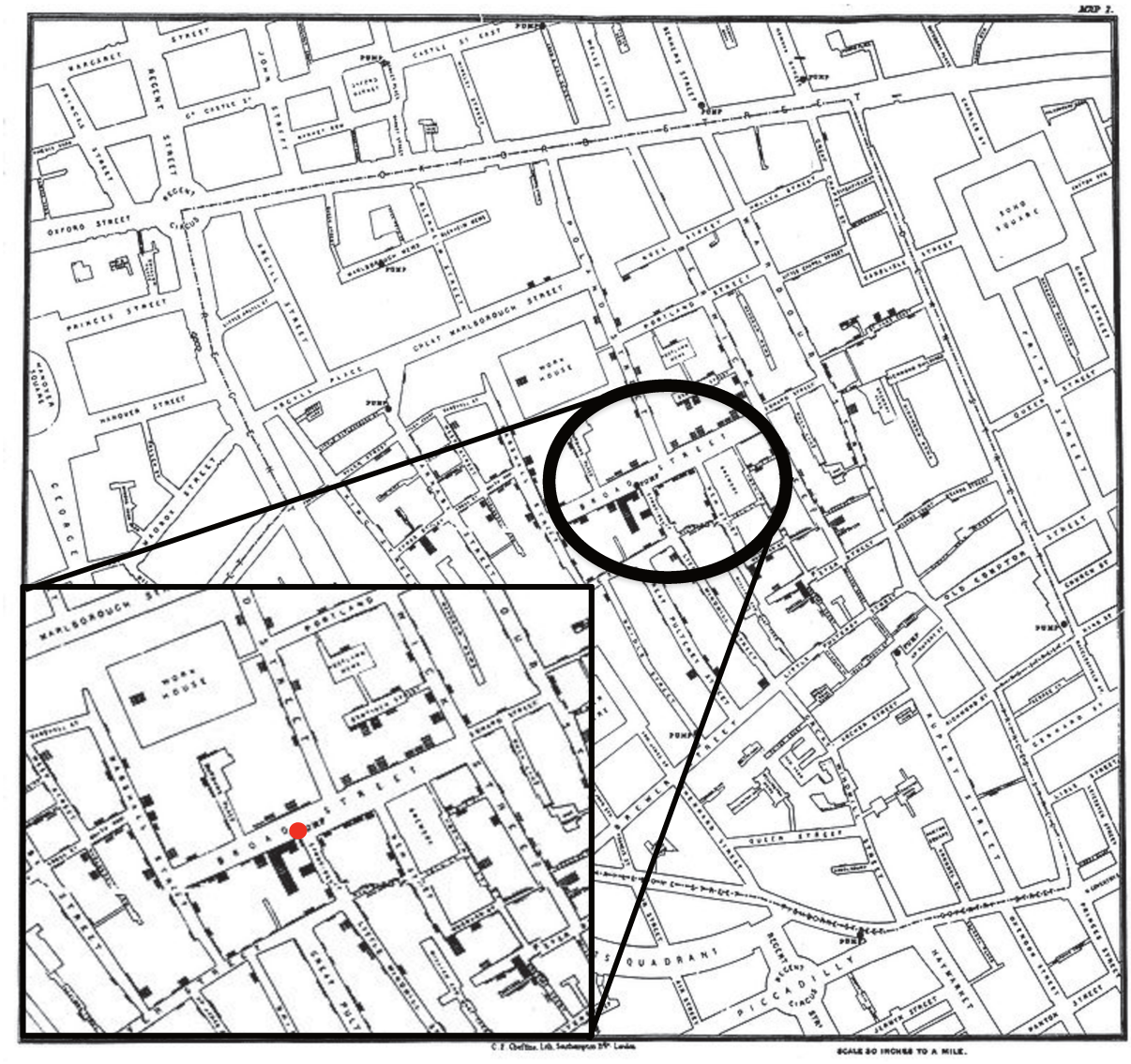


Figure 1 : Localisation des cas de choléra dans le quartier de Soho, Londres dans l'étude de John Snow en 1855. Dans le cercle noir, la rue de Broad Street. Les rectangles noirs représentent le nombre de cas à chaque adresse, en rouge l'emplacement de la pompe de Broad Street.

Les outils utilisés par John Snow restent simples : il y a peu de données et seule une analyse descriptive est réalisée sans utilisation de statistiques inférentielles ni ajustement des analyses *a minima* sur la taille de la population. Il s'agissait des prémices de ce qu'est devenue l'analyse spatiale moderne.

Cette dernière a été permise par l'augmentation de la disponibilité de base de données de grande dimension, le développement du *data reuse* et celui du géoréférencement. En effet, l'intégration de systèmes de positionnement par satellites dans de nombreux systèmes (smartphones, tablettes, montres connectées, véhicules…) permettent de décrire précisément dans l'espace les trajets quotidiens, les lieux de travail et de vie, les modes de transports… Sur le plan de la santé, il est devenu possible de décrire la consommation de soins, de connaitre l'épidémiologie de pathologies et d'analyser leurs déterminants écologiques. Ces analyses sont d'autant plus complexes que le volume de données est important et qu'il est nécessaire de limiter l'influence de biais méthodologiques dans le contexte de *l'Evidence Based Medicine*.

L'informatique est indispensable pour permettre la récolte et le stockage des données, leur traitement ainsi que leur analyse. Le regain d'intérêt de l'analyse spatiale en santé résulte de la mise à disposition de ces outils plus ou moins spécifiques :

- Système de récolte et de stockage de données : Le développement du *cloud*, et de l'*open data* permettent de récupérer des données disponibles sur des serveurs y compris publics.
- Manipulation des données : Un Système d'Information Géographique (SIG, *Geographic information system, GIS* en anglais) permettant de manipuler les données spatiales, de réaliser des fonds de cartes, les notions de systèmes de projections, de distances et de voisins… Les plus connus sont ArcGIS® et QGIS (16,17). Ces

logiciels peuvent aussi être complétés par des outils statistiques tels que R. Il s'agit en effet d'un exemple de logiciel performant et gratuit qui ne cesse d'évoluer du fait de la mise à disposition d'un grand nombre de *packages* (18).

- Analyse des données et affichages : des logiciels généralistes tels que R peuvent suffire. Néanmoins, dans certaines situations, il faudra faire appel aux SIG permettant de réaliser une partie des analyses et de générer un affichage (19). Enfin, des outils très spécifiques seront nécessaires pour la réalisation de certaines analyses spatiales tels que le logiciel Winbugs® pour les régressions écologiques bayésiennes ou le logiciel Satscan® pour la recherche d'agrégats d'évènements (20,21)

Le conjonction de la disponibilité des outils et de la mise à disposition de nombreuses bases de grandes dimensions a ouvert de nouveaux champs en épidémiologie via l'analyse spatiale (22). L'application en santé des concepts d'analyse spatiale a dû être développée. La première étape a été celle de définir les données de santé manipulées selon leur type spatial.

B. Données manipulées

En analyse spatiale, les données rencontrées peuvent être catégorisées en trois grands types. Elles ne peuvent être croisées les unes avec les autres sans avoir bien définis leurs propriétés. Il est par ailleurs nécessaire de les manipuler avec l'aide d'outils adaptés. On rencontre ainsi les données ponctuelles, les données géostatistiques et les données latticielles pour lesquelles seront données systématiquement des exemples en santé.

i. Données ponctuelles

Les données ponctuelles se définissent par des points localisés par des coordonnées géographiques. Leur nombre ainsi que leur distribution dans l'espace sont aléatoires. On parle

ici de processus ponctuels. Dans cette situation, la localisation de l'objet étudié est l'objet de l'analyse. Un exemple habituel est celui d'une forêt dans laquelle on s'intéresserait à la position de chaque arbre.

Dans le domaine de la santé, les données ponctuelles pourraient être par exemple des coordonnées de lieu de résidence des patients ou d'emplacements d'industries polluantes. On peut aussi imaginer recenser l'ensemble des adresses des interventions du SAMU pour arrêt cardiaque. Il serait ainsi possible d'en analyser la répartition spatiale.

ii. Données géostatistiques

Les données géostatistiques se définissent par des points localisés par des coordonnées géographiques dont le nombre ainsi que la répartition sur le territoire sont choisis et fixes. Chaque site géographique se distingue par ses coordonnées géographiques mais aussi par les valeurs des mesures qui y sont effectuées. En d'autres termes, la position est fixe alors que la mesure est modélisée par une variable aléatoire. Les données géostatistiques peuvent provenir par exemple de sites de captage de mesure de la qualité des nappes phréatiques. En effet dans ce cas, le site de captage a une position fixe définie par l'observateur alors que la qualité de l'eau, par exemple la contamination en *Escherichia Coli*, est aléatoire.

Il est possible grâce aux méthodes statistiques d'interpolation spatiale d'estimer les valeurs sur l'ensemble du territoire à partir des mesures de chaque site (23). Ces méthodes permettent à partir de données connues en un certain nombre de points fixes d'estimer les valeurs d'une variable en tout point d'une zone géographique. La précision de l'interpolation sera améliorée d'une part par l'augmentation du nombre de données initiales et d'autres part par l'augmentation de l'homogénéité de la répartition des points observés. On peut citer comme exemple fréquent celui de mesure de la pluviométrie en plusieurs points d'un territoire puis par interpolation spatiale, la connaissance de la valeur de pluviométrie sur l'ensemble de la zone étudiée (24).

Il est utile en santé par exemple de chercher des liens entre la pollution de l'air et certains évènement de santé comme cela a pu être fait pour les faibles poids de naissance (25). Dans ce contexte, les données de pollution proviennent de mesure de la pollution aérienne en des points précis. Les stations de mesure de polluants dans l'air sont situées en différentes localisations du territoire d'une région mais leur emplacement est fixé par l'association agréée de surveillance régionale de la qualité de l'air telle qu'Atmo dans les Hauts de France (26). Ces dernières ont en effet pour rôles de « surveiller l'air, informer et alerter en cas de phénomènes de pollution atmosphérique ».

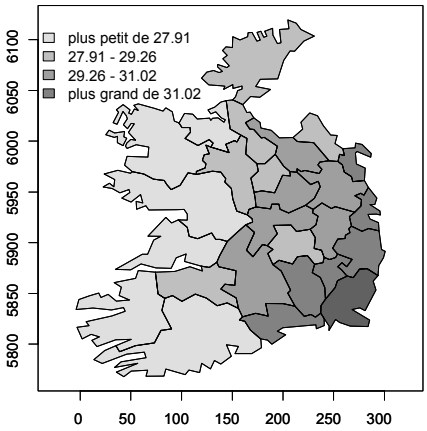
iii. Données latticielles

Les données latticielles sont celles qui sont le plus souvent manipulées en santé. Elles sont appelées couramment données agrégées. Elles sont fréquentes dans les bases de données médico-administratives car elles présentent deux avantages :

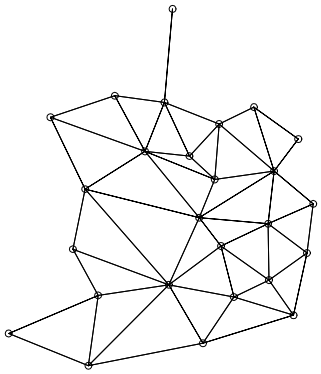
- Facilité de récupération : un nom de ville ou un code postal suffisent ;
- Anonymisation : en agrégeant les données, elles entrainent une anonymisation permettant de conserver le secret médical.

Ce type de données se définie par une variable aléatoire indexée par un ensemble discret défini comme un réseau discret structuré par un graphe de voisinage. L'ensemble des variables aléatoires constituent un champ aléatoire dont l'espace d'état peut être de nature variée (entier naturel, nombre réél, etc). En d'autres termes il s'agit de données concernant des sous-ensembles de la population localisés en plusieurs points fixes de l'espace. Un exemple de données latticielles est données par la Figure 2 repris du travail de thèse de Genin (27). La variable aléatoire est ici le pourcentage de la population présentant le groupe sanguin A ramené en chaque point du graphe de voisinage correspondant à un comté d'Irlande.

Figure 2 : (a) Pourcentage de la population présentant le groupe sanguin A dans les 26 comtés d'Irlande. (b) Graphe de voisinage des 26 comtés.



(a)



(b)

Dans ce type de données, le niveau d'agrégation est un paramètre important. Ainsi pour reprendre les données de l'exemple en Figure 2, les populations de groupe sanguin A sont représentées au comté mais pourraient être représentées au pays ou à la commune. La représentation de la population de groupe sanguin A à ces différents niveaux modifie la perception des résultats. Supposons que l'on souhaite par exemple répartir sur le territoire les stocks de concentrés de globules rouges (CGR) de groupe A sang en fonction de la population de groupe A, en l'absence d'informations plus précises que le pourcentage de population de groupe sanguin A au pays, la répartition de CGR de groupe A devra être homogène sur le territoire de l'Irlande. Avec des données au niveau du comté, il semble nécessaire d'augmenter la proportions de CGR de groupe A dans les comtés de l'est de l'Irlande par rapport aux comtés de l'ouest. Cependant, la répartition devra être homogène dans le comté. Une connaissance des données au niveau de la commune permettra une répartition des CGR encore mieux adaptée aux besoins de la population.

Les données de ce type sont fréquentes dans la littérature. Il s'agit par exemple du nombre de cancers inflammatoires du sein par *county* aux USA utilisé dans l'étude de Scott et al., du

nombre de patients admis pour pathologie psychiatrique par *prefecture* aux Japon utilisé dans l'étude de Takahashi et al. ou encore du nombre de cas de Chorée de Huntington par *district* en Espagne dans l'étude de Sanchez-Diaz et al. (28–30).

Les outils informatiques tels que QGIS ou R vont permettre de pouvoir manipuler et analyser ces différentes données. En analyse spatiale en santé, il est particulièrement important de pouvoir disposer de méthodes efficaces d'analyses des données latticielles compte tenu de leur fréquence. Ces analyses ne sont pas une finalité et restent un moyen de répondre aux hypothèses posées face à une pathologie. Selon l'hypothèse, il va être nécessaire de choisir la méthode d'analyse spatiale adaptée.

III. Méthodes d'analyse spatiale pour données latticielles

Les méthodes d'analyses spatiales qui seront vues dans ce chapitre seront celles s'appliquant sur les données latticielles. En effet, comme il a été vu ci avant, elles sont particulièrement rencontrées en santé. Les notions de répartition spatiale, de dépendance liée à la proximité spatiale et de facteurs de risque écologiques sont autant d'éléments qui peuvent être pris en compte lors des analyses. La finalité du traitement de données est bien de pouvoir expérimenter et tester des hypothèses scientifiques. John Snow avait par exemple dès le départ posé l'hypothèse d'une contamination par l'eau pour les épidémies de choléra. Ainsi les différentes méthodes d'analyses spatiales ne sont pas exclusives les unes des autres et se complètent pour permettre de tester les différentes hypothèses. Pour chaque type d'analyse spatiale, il sera présenté un exemple de méthode correspondante. Tout d'abord la représentation cartographique des maladies (*disease mapping*) sera abordée avec le calcul du *standardized incidence ratio* (SIR), puis la recherche d'agrégat sera étudiée par les

statistiques de scan spatiales de Kulldorff et enfin la régression écologique sera présentée à travers l'utilisation du modèle bayésien de Besag, York et Mollié.

A. Représentation cartographique des maladies

La première étape d'une analyse spatiale est de connaître la répartition spatiale d'une maladie. La lecture d'une telle carte peut, par exemple, permettre de générer des hypothèses sur les liens entre évènements de santé et niveau de revenu des personnes ou avec l'âge de la population dans une région donnée. Il est tout à fait possible de ne représenter que le nombre de cas d'une pathologie sur la carte mais ce nombre sera dépendant de la taille de la population. La première étape sera d'utiliser des taux d'évènements rapportés à la population. Cependant si l'on compare les taux d'incidences brutes de maladies d'Alzheimer, les zones se démarquant par des taux élevés seront celles ayant probablement une population plus âgée. Il est nécessaire de pouvoir ajuster sur les facteurs de confusion important tels que l'âge et le sexe. Cet ajustement peut se faire par méthode directe ou indirecte, la méthode directe nécessitant une population de référence, l'autre non. L'ajustement permet de calculer un nombre de cas attendus E par unité spatiale (31,32). Cette valeur E repose sur l'hypothèse d'une incidence homogène de la maladie sur l'ensemble de la zone étudiée. Cette variable E est ensuite comparée au nombre de cas observés N par unité spatiale. L'indicateur est appelé rapport standardisé d'incidence (*standardized incidence ratio, SIR*) qui est une estimation du risque relatif de l'unité spatiale considérée par rapport à la moyenne de la zone étudiée. Considérons une unité spatiale *i*, le SIR se calcule par l'expression suivante :

$$SIR_i = \frac{N_i}{E_i}$$

où N_i désigne le nombre de cas observés au sein de l'unité spatiale et E_i correspond au nombre de cas attendus issu d'une méthode de standardisation.

Un SIR supérieur à 1 correspondra à un risque relatif supérieur à 1 et signalera une zone géographique présentant une sur-incidence par rapport à l'ensemble de la zone étudiée et un SIR inférieur à 1 correspondra à un risque relatif inférieur à 1 et signalera une zone géographique avec une sous-incidence de la pathologie par rapport à l'ensemble de la zone étudiée.

Par ailleurs, les SIR permettent d'ajuster des taux d'incidence sur des facteurs de confusions. Cette standardisation nécessite des données de populations stratifiées par variable d'ajustement. Ainsi elle est habituellement faite sur l'âge et/ou le sexe car ce sont des données facilement disponibles auprès des instituts de recensement tels que l'Insee.

Cependant, les SIR sont performants sous conditions de population suffisante et d'évènements suffisamment fréquents. Les unités spatiales ne respectant pas ces conditions peuvent amener à des calculs de SIR aboutissant à des valeurs extrêmes concluant faussement à une zone de sur ou de sous risque importante. Afin de palier ces limites, il existe une technique de lissage des SIR. Ce lissage permet de prendre en compte l'autocorrélation spatiale, c'est à dire le fait que des unités géographiquement proches interagissent entre elles et ont une probabilité supérieure de se ressembler. Cette méthode a par exemple été utilisée dans une étude sur les cas de maladie de Crohn à l'échelle cantonale dans le Nord Pas de Calais dont les résultats sont présentés dans la Figure 3 (33). La méthode bayésienne permettant ce lissage sera vue au paragraphe III.C.

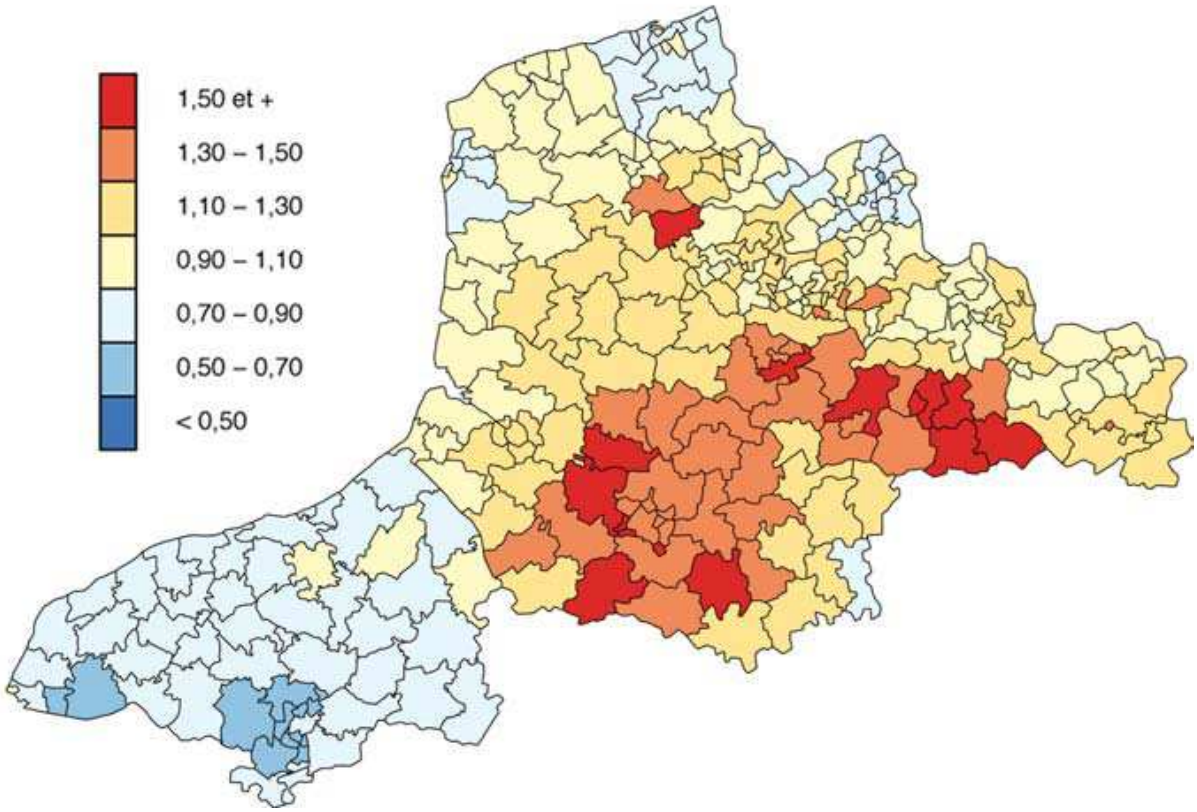


Figure 3 : Carte des SIR lissés dans l'étude de la maladie de Crohn à l'échelle cantonale à partir du registre Epimad pour la région Nord-Pas de Calais.

Enfin les SIR permettent de mettre en évidence une hétérogénéité spatiale mais sont purement descriptifs et ne permettent ni d'interprétation clinique quant aux liens avec d'éventuels facteurs de risque ni de mise en évidence d'agrégats. Cette recherche objective d'agrégats statistiquement significatifs doit être faite par l'utilisation de méthodes d'analyse spatiale spécifiques.

B. Détection de cluster

Après avoir cartographié une maladie, il est intéressant de rechercher un agrégat (cluster) de cas de zones géographiques se démarquant par une sur-incidence ou une sous-incidence statistiquement significative. John Snow a réalisé cette démarche dans son étude sur le choléra à Londres. Néanmoins sa recherche d'agrégat était subjective et aucune probabilité critique

n'a été calculée. Il existe actuellement des méthodes objectives de recherche d'agrégats, c'est à dire sans connaissances *a priori* sur la localisation et la taille des clusters recherchés.

La méthode de statistique de scan spatiale de Kulldorff est une des plus fréquentes (une recherche sur Google Scholar et PubMed avec les termes « kulldorff spatial scan statistic » renvoie par exemple respectivement 5020 et 381 résultats) car c'est une des plus puissantes méthodes de détection de cluster à ce jour, relativement simple d'utilisation et mise à disposition dans le logiciel gratuit SatScan (20,34–36). L'objectif de cette méthode est de détecter des agrégats spatiaux statistiquement significatifs sans biais de présélection.

La statistique de scan est une variable de décision aléatoire d'un test statistique dont les hypothèses sont les suivantes :

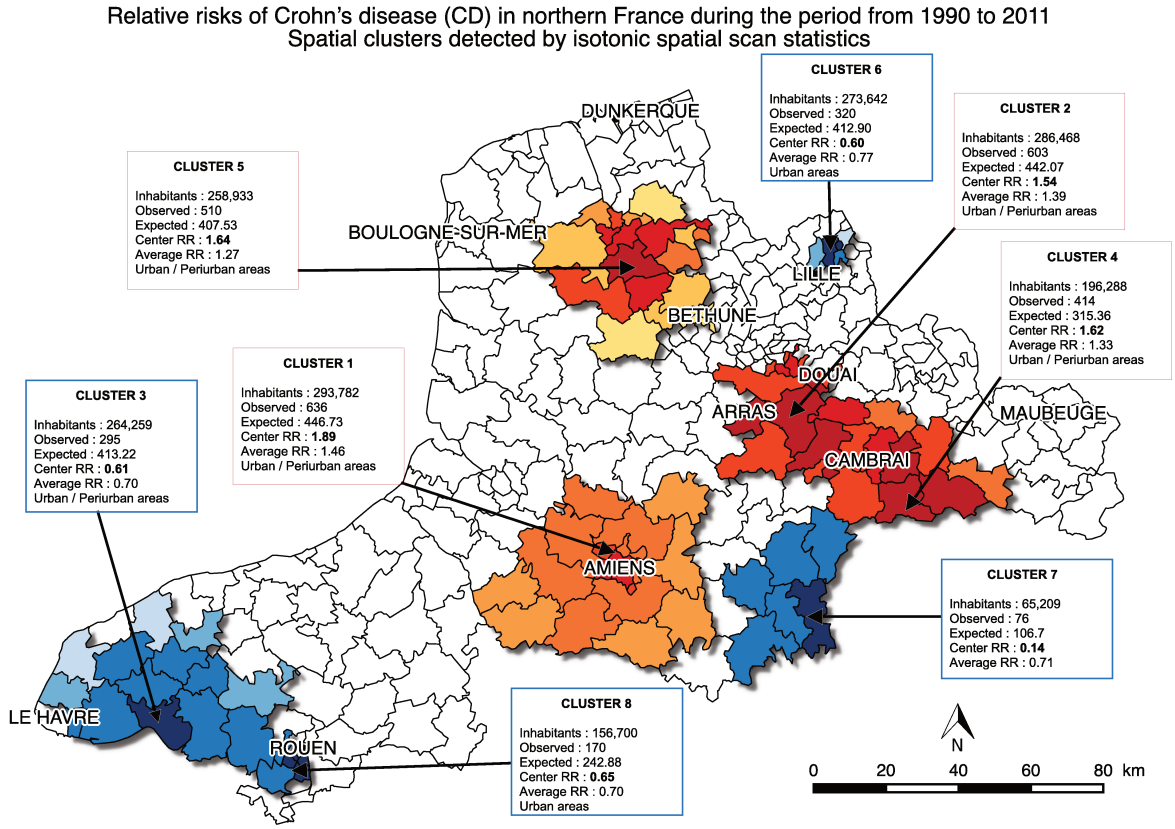
- Hypothèse H0 : La probabilité d'apparition d'un événement est constante sur l'ensemble de la zone étudiée.
- Hypothèse H1 : Il existe une sous-zone (cluster) dans laquelle la probabilité d'apparition d'un évènement est supérieure à celle de la zone étudiée.

La méthode est constituée de 2 étapes effectuées par le logiciel SatScan : une phase de détection et une phase d'inférence statistique. Le lecteur pourra se rapporter au travail publié en 1997 de Kulldorff pour une présentation complète du modèle (37). Le cluster le plus probable, le *most likely cluster,* sera d'abord mis en évidence. Kulldorf a proposé une méthode permettant d'estimer les probabilités critiques des clusters secondaires tenant compte de l'existence du *most likely cluster (38)*. Les clusters peuvent être de forme circulaire ou elliptique et de tailles variables (dont la taille maximum est paramétrée par l'utilisateur). Il est par ailleurs possible de réaliser un ajustement sur des covariables telles que l'âge et le sexe. Enfin, lorsqu'un cluster est grand, l'interprétation épidémiologique devient difficile car la population sous jacente concernée est grande. Il est alors intéressant d'utiliser la méthode de

statistiques de scan isotoniques qui consiste à modéliser une fonction de risque décroissante du centre vers la périphérie des clusters. Il est ainsi mis en évidence un « épicentre » du cluster, c'est à dire la zone du cluster avec le risque relatif le plus éloigné de 1.

Cette méthode a été appliquée dans la recherche de cluster d'incidence de cas de maladie de Crohn sur les départements du Nord, du Pas-de-Calais, de la Somme et de la Seine maritime à partir des données de cas du registre Epimad et des données de populations de l'Insee. Cela a permis de mettre en évidence 4 clusters de sur-incidence et 4 clusters de sous-incidence présentés sur la Figure 4. Dans un second temps, les clusters ont été caractérisé par des paramètres écologiques tels que leur démographie, leur tissu industriel ou le type de surface agricole.

Figure 4 : Cluster isotonique de maladie de Crohn à partir des données du registre Epimad dans le Nord de la France



La méthode de statistique de scan spatiale de Kulldorff, bien qu'ayant des limites en termes de temps de calcul et de forme des agrégats recherchés est particulièrement intéressante pour mettre en évidence des clusters de sur- et de sous-incidence. De plus, il existe des extensions de cette méthode qui ajoutent à la composante spatiale une composante temporelle. Kulldorff les a appliquées sur des données relatives au cancer cérébraux au Nouveau Mexique aux USA (39).

Néanmoins elle ne permet pas de mettre en lien directement les données de santé avec des variables écologiques. Les unités spatiales doivent en effet être caractérisées dans un second temps et comparées selon leur appartenance ou non à un cluster. C'est pourquoi il est nécessaire d'utiliser dans ce cas la régression écologique, outil des analyses spatiales en santé permettant d'identifier des facteurs de risques ou des facteurs protecteurs écologiques.

C. Régression écologique : le modèle de Besag York et Mollié

La modèle de Besag, York et Mollié (BYM) a initialement été développé pour réaliser de la restauration d'image mais il est aujourd'hui largement utilisé dans le domaine de l'analyse spatiale en santé pour calculer des régressions écologiques (40,41). Il s'agit d'un modèle bayésien hiérarchique qui découle de la nécessité de prendre en compte deux éléments, modélisés par des effets aléatoires :

- Une variabilité sans motif spatial, appelée hétérogénéité non structurée, qui ne dépend pas de l'unité spatiale considérée ;
- Une variabilité avec dépendance spatiale, aussi appelée hétérogénéité structurée. Cette variabilité rend compte de la corrélation entre voisins.

La corrélation spatiale signifie que plus des unités spatiales sont voisines, plus elles sont interdépendantes. C'est le concept imagé par Tobler dans sa 1ère loi de la géographie : « Toutes choses sont reliées à tous le reste, mais les choses proches sont plus liées que celles

qui sont éloignées » (42). Le lecteur pourra se rapporter à l'ouvrage d'Andrew B. Lawson pour une présentation complète du modèle (43).

La première utilisation du modèle BYM est de lisser des SIR comme présenté au paragraphe III.A dans la Figure 3. Le calcul des SIR n'est plus indépendant pour chaque unité spatiale mais prend systématiquement en compte les voisins.
La seconde utilisation du modèle BYM est sa capacité à calculer des régressions écologiques. La réalisation de modèles multivariés permet d'une part des ajustements sur les facteurs de confusion et d'autre part d'isoler pour chaque covariable son lien « réel » avec la probabilité de survenue de l'évènement. Chaque covariable du modèle sera caractérisée par un risque relatif ainsi qu'un intervalle de crédibilité. L'évaluation des modèles se fait à travers le *Deviation index criterion* (DIC), semblable à l'*Akaike index criterion* (AIC) utilisé dans les modèle de statistiques fréquentistes. Ce modèle a par exemple été utilisé par Rooney et al. pour la recherche d'un lien entre sclérose latérale amyotrophique et les constituants du sol (44). Il est par ailleurs possible d'étendre le modèle en ajoutant à la composante spatiale une composante temporelle (45).

Après l'application de ces différentes méthodes d'analyses spatiales, il est possible de connaître de façon approfondie la répartition d'une maladie, l'existence d'une hétérogénéité spatiale sous forme de cluster ainsi que les facteurs écologiques prédictifs de la maladie. Cependant, ces méthodes supposent de pouvoir utiliser conjointement les bases de données médicales et les bases de données écologiques. Dans le cas contraire, il s'agit du problème déjà décrit dans la littérature sous le nom de *change of support problem.*

IV. Le Change of support problem

A. Définition

Le *change of support problem* (*COSP*), se définit par l'absence d'identifiant spatial unique permettant la fusion de 2 bases de données. En effet, les méthodes d'analyses spatiales reposent systématiquement sur le croisement des bases de données médicales et des bases de données écologiques. A titre d'exemple, les données de cas sont *a minima* ajustées sur la taille de la population afin de calculer soit une prévalence soit une incidence. La fusion des 2 bases de données est réalisée au moyen d'un identifiant spatial unique. Il s'agit de la situation type présenté en Figure 5. Gotway et Young ont décrit les différentes situations possibles ne permettant pas la fusion sous le terme de *COSP*, répertoriées au sein du Tableau 2 (46). D'autres exemples particulièrement rencontrés en santé sont présentés dans le Tableau 3.

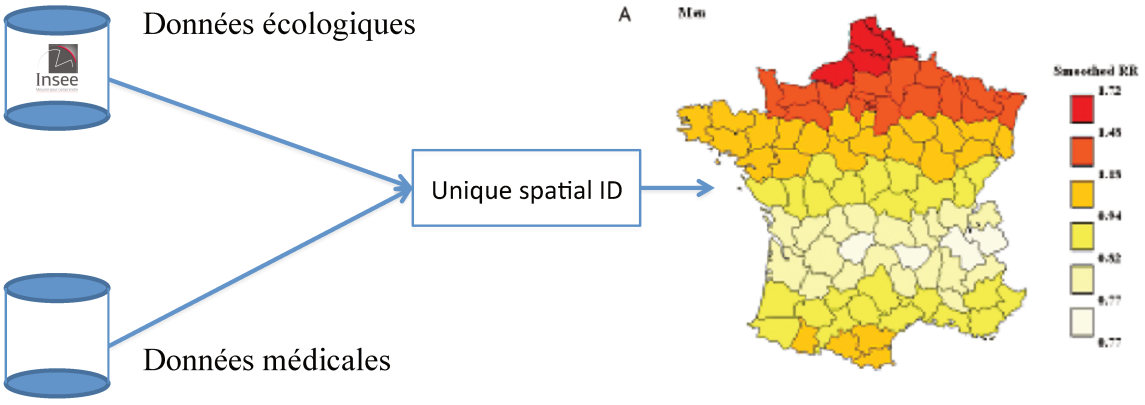


Figure 5 : Schéma conceptuel classique dans la réalisation d'analyses spatiales en santé à partir de base de données écologiques et de bases de données médicales. La carte représente les SIR lissés de maladie de Crohn issu d'une étude de Nerich et al. (47).

Tableau 2 : Les différentes situations du *change of support problem* selon Gotway et Young.

Table 1. Examples of COSPs

We observe or analyze	*But the nature of the process is*	*Examples*
Point	Point	Point kriging; prediction of undersampled variables
Area	Point	Ecological inference; quadrat counts
Point	Line	Contouring
Point	Area	Use of areal centroids; spatial smoothing; block kriging
Area	Area	The MAUP; areal interpolation; incompatible/misaligned zones
Point	Surface	Trend surface analysis; environmental monitoring; exposure assessment
Area	Surface	Remote sensing; multiresolution images; image analysis

Tableau 3 : Exemple de situations rencontrées en santé qui relèvent du COSP.

Nature du COSP	Exemples en santé
Zones - points	Patients agrégés à la commune – industrie polluantes
Points - Lignes	Coordonnées de cas – Emplacement d'une route
Zone - Zone	Patients agrégés aux codes postaux – Données socio-économiques à la commune

Le COSP peut être catégorisé en deux types de situations ne permettant pas la fusion des 2 bases de données :

- Les identifiants spatiaux correspondent à des données de types différents. Il s'agit par exemple de situation de fusion entre des données de cas agrégées à la commune et des données ponctuelles de pollution.

- Les identifiants spatiaux sont de même type mais correspondent à des zones géographiques de tailles différentes. Il s'agit par exemple de la situation de croisement entre des données de cas agrégées au code postal et des données socio-économiques agrégées à la commune.

La situation étudiée dans ce travail est particulièrement celle des *area to area* aussi appelée *misaligned zones*. Les codes postaux constituent fréquemment l'unité spatiale des bases médico-administratives car étant à la fois simple à récupérer et permettant un bon degré d'anonymisation des données. C'est le cas du Programme médicalisé des systèmes d'information (PMSI) mais aussi des bases du National Health System (NHS) au Royaume-Uni ou de la base Statewide Planning and Research Cooperative System (SPARCS) de l'Etat de New-York (48–50).

Les données socio-économiques nécessaires au calcul des taux d'incidence ou de prévalence sont produites par les instituts de recensement nationaux. Ces instituts s'appuient sur le découpage administratif et en particulier celui des communes pour produire les données comme par exemple celles issues de l'état civil (population, naissances, décès).

Afin de pouvoir exploiter ces données riches d'informations, il est nécessaire de trouver des solutions au COSP (51,52).

B. Solutions au change of support problem

La problématique des *misaligned zones* trouve actuellement deux solutions possibles :

- L'agrégation à un niveau spatial supérieur ;
- L'utilisation de méthodes de désagrégation spatiale.

L'agrégation à un niveau supérieur consiste à transformer des données disponibles à une unité spatiale A en données disponibles à unité spatiale B de taille supérieure. Elle présente l'avantage de ne pas modifier les valeurs des données. Ce sont le plus souvent des unités administratives qui sont utilisées dans la base finale d'analyse spatiale. La transformation de données disponibles à la commune ou au code postal en données disponibles au département est un exemple dans le cas français. Le problème de l'agrégation est la perte de précision dans les analyses avec une mise en évidence moindre de phénomènes locaux et une augmentation du biais écologique (2,9,53). C'est la solution retenue par l'Irdes dans une étude sur les pratiques chirurgicales qui présente des résultats au niveau départemental (54).

Les méthodes de désagrégation spatiale sont variées. Elles consistent toujours à transformer des données d'une unité spatiale en une autre unité spatiale d'échelle plus fine. Ils existent les méthodes suivantes :

- L'imputation aléatoire des données. Dans cette situation, les données de larges unités spatiales sont reparties aléatoirement dans les unités spatiales plus fines incluses dans l'unité spatiale initiale. C'est une solution retenue par la Drees dans une étude sur les distances d'hospitalisation (55).
- L'imputation selon des modèles plus ou moins complexes. Dans ce cas, l'imputation n'est plus aléatoire mais est réalisée en ajustant sur des paramètres tels que la surface, la population, le degré d'urbanisation…
- L'interpolation spatiale. Cette méthode est spécifique aux statistiques spatiales. Elle permet de prendre en compte les valeurs des unités spatiales sur l'ensemble de la zone et d'en estimer les valeurs aux unités spatiale d'échelles plus fines. C'est la méthode la plus fine mais la plus complexe. Un des exemples de cette méthode est le kriegage (46).

Les méthodes de désagrégation spatiale ont toutes pour inconvénient d'introduire des erreurs dans l'estimation des variables aux unités spatiales d'échelles plus fines. Ces erreurs nuisent à la qualité de l'analyse spatiale finale et sont difficilement mesurables.

Le *change of support problem* reste un obstacle important à l'utilisation des grandes bases de données médico-administratives en analyse spatiale, champ pourtant riche des possibilités qu'offre le croisement avec les bases de données socio-économiques et environnementales très largement mises à disposition aujourd'hui.

V. Données spatiales utilisées en santé

Les données de santé accessibles aux chercheurs sont nombreuses. Il existe depuis longtemps des registres de maladies tels que le registre Epimad pour les maladies inflammatoires chroniques de l'intestin ou le registre REIN de l'insuffisance rénale. Ces bases de données conçues pour l'épidémiologie sont complétées par les données médico-administratives que sont la base du PMSI ou celle du Système National des Données de Santé (SNDS).

A. Données de santé

Les données de santé initialement utilisées en analyse spatiale en santé étaient celles des registres. En effet, ces bases de données sont conçues dans un objectif épidémiologique. Leur exhaustivité est régulièrement évaluée, les facteurs de confusion potentiels y sont le plus souvent rapportés et les autorisations de recherche permettent une localisation précise des événements.

Les données médico-administratives rentrent par contre dans le cadre du *data reuse*. Cette pratique consiste à réutiliser en recherche des bases de données constituées pour d'autres objectifs. Le PMSI est par exemple la base de données des séjours hospitaliers français. Les

données d'activités qui en sont issues servent avant tout au financement des établissements de santé dans le cadre de la Tarification à l'Activité. Les données disponibles dans le PMSI sont variées. Cette base contient des informations administratives telles que l'âge, le sexe, le lieu de résidence sous une forme proche du code postal, la durée de séjour ou l'établissement de santé (identifié par le numéro FINESS) ainsi que des informations médicales telles que le diagnostic principal du séjour correspondant au motif d'hospitalisation depuis 2009 et les diagnostics associés correspondant aux comorbidités du patient enregistrés sous forme de code de la 10ème révision de la classification international des maladies (CIM-10) ou les actes médicaux pratiqués enregistrés sous forme d'un code de la classification commune des actes médicaux (CCAM). La base PMSI est administrée par l'Agence technique de l'information hospitalière (ATIH) qui la met à disposition des équipes de recherche sous couvert d'une autorisation de la Commission nationale de l'informatique et des libertés (CNIL).

L'utilisation en épidémiologie du PMSI est documentée. La qualité des données est aujourd'hui suffisante pour étudier plusieurs pathologies telles que les cancers ou la périnatalité malgré certaines limites qui persistent tels que les variations de pratique de codage dans le temps ou selon les établissements (56–60). Le nombre important de séjours, l'étendue géographique couverte et la période de temps qui y est intégrée en fait un outil privilégié pour analyser des évènements de santé variés qui peuvent être non spécifiques (taux de mortalité, ré-hospitalisations) ou spécifiques (calcul de taux de prévalence ou d'incidence de pathologies). L'utilisation du PMSI dans l'analyse spatiale est plus récente que son utilisation en épidémiologie classique. En effet, l'unité spatiale utilisée dans la base de données est un code géographique basé sur les codes postaux adaptés par l'ATIH. Cela rend plus complexe son utilisation. Néanmoins, cela existe comme en atteste les études de la Drees et de l'Irdes ou de différents Observatoire Régionaux de la Santé (ORS) (54,55,61).

Dans chacun de ces cas, les données du PMSI disponibles aux codes géographiques ont subi une agrégation pour être disponibles à des unités administratives plus grossières (département) ou ont été traitées par des méthodes de désagrégation spatiale pour être analysées à des échelles plus fines (communes). A notre connaissance, seule une étude française sur les accidents vasculaires cérébraux a utilisé l'unité spatiale du code géographique PMSI pour réaliser une analyse spatiale (62).

Par ailleurs, le SNDS permet d'avoir accès à la base de données du Système National d'Information Inter Régime de l'Assurance Maladie (SNIIRAM). Cette base de données intègre toutes les données concernant les assurés sur le plan administratif (lieu de résidence à la commune, sexe, âge..) et médical (affection longues durées, consommation de soins en ville…) (63). La base est anonyme mais il est possible de chaîner les patients avec la base PMSI. De plus, il existe un extrait plus accessible appelé Echantillon Généraliste des Bénéficiaires (EGB) représentant 1/97ème de la base. Néanmoins, les qualités épidémiologiques de cette base d'utilisation récente et d'accès plus restreint sont moins connues. Cependant le SNIIRAM et l'EGB semblent riches d'opportunités (64).

En conclusion, les données médicales utilisables en analyse spatiales en santé sont nombreuses. En particulier, l'utilisation des différentes bases médico-administratives en épidémiologie se développe rapidement. Un des principaux points de l'analyse spatiale est de pouvoir croiser ces bases de données médicales avec différentes bases de données écologiques.

B. Les bases de données écologiques

Les données écologiques disponibles en France sont très nombreuses et cette présentation succincte ne saurait être exhaustive. Cependant, seront d'abord présentées les données mises à

dispositions par l'INSEE, puis des données d'occupation des sols et enfin différentes données relatives à la pollution.

Tout d'abord, les données écologiques les plus accessibles sont celles mise à disposition par l'Insee. Elles couvrent des domaines variés tels que la démographie, les données socio-économiques ou encore le tissu industriel. Ces bases de données sont publiques et disponibles pour de longues périodes de temps. Les données sont fréquemment agrégées à l'échelle de l'unité spatiale communale dont l'identifiant spatial est le code géographique Insee. Les données socio économiques permettent par exemple de calculer des indices de défaveur sociale telle que l'indice de Townsend ou le French EDI (*European deprivation index*, indice européen de défaveur sociale) (65,5). Il est aussi possible d'analyser directement certains éléments telles que la structure d'âge, le niveau de revenu ou encore la structure de catégories socio-professionnelles.

Ensuite sur le plan environnemental, il existe la base Corine Land Cover qui permet d'avoir accès au type d'occupation du sol pour chaque commune. Cette base de données européenne est coordonnée en France par le ministère de l'environnement. Il en existe plusieurs nomenclatures allant de 5 catégories différentes au premier niveau jusque 44 catégories au 3ème niveau. Cela permet de différencier par exemple un tissu urbain continu, une emprise routière ou ferroviaire, une décharge ou encore une terre arable non irriguée. Les données sont présentées sous la forme de surface par commune.

Enfin les données relatives à la pollution sont d'un grand intérêt tant les liens entre différentes pollutions atmosphériques, aquatiques ou des sols et impact sur la santé sont soupçonnés (25,66–70). Les données nationales sont encore relativement rares mais les bases de données

de sites polluants et de sites et sols pollués sont d'importantes sources d'informations accessibles publiquement. L'inventaire historique des sites industriels et activités en service (BASIAS) et l'inventaire des sites et sols pollués (BASOL) contiennent l'ensemble des sites polluants, potentiellement polluants et pollués localisés par des coordonnées géographiques ou une adresse et caractérisés par l'activité exercée. Cela rend par exemple possible la recherche d'une proximité entre les événements de santé et les sites polluants. Ces données nécessitent un important *data management* avant réutilisation, en particulier *via* un géocodage (détermination des coordonnées géographiques d'une adresse) indispensable pour les sites sans coordonnées géographiques. L'autre source française de données de pollution est l'ensemble constitué par les réseaux de surveillance de la qualité de l'air rendus obligatoires par la Loi n° 96-1236 du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie (71). L'air est ainsi surveillé attentivement que cela soit sur certains polluants spécifiques tels que les oxydes d'azote ou plus largement sur les microparticules. L'ensemble du territoire français est couvert de façon plus ou moins précise. L'utilisation de ces données nécessite l'usage de techniques spécifiques permettant de les rendre interopérables avec les données agrégées classiquement rencontrées en santé.

Ces nombreuses bases de données socio-économiques et environnementales sont riches d'informations permettant la recherche de facteurs de risque de pathologies variées. Cependant, les données médicales et les données écologiques ne partagent pas systématiquement des identifiants spatiaux communs. C'est par exemple le cas du PMSI qui contient un identifiant spatial spécifique issu du code postal. Cette problématique d'unité spatiale ne se recoupant pas s'inscrit dans la situation des *misaligned zone* décrites dans le Tableau 2 reprenant la classification des COSP. C'est pourquoi il a semblé important de surmonter ce problème sans dégrader l'analyse spatiale.

Problématique générale

Les données disponibles pour l'analyse spatiale en santé sont très nombreuses. Il existe des bases de données médicales de grandes dimensions qu'il est particulièrement intéressant de pouvoir croiser avec des bases de données écologiques socio-économiques et environnementales. Le potentiel des hypothèses pouvant être testées et des analyses pouvant être réalisées est majeur. La recherche d'agrégat peut permettre de cibler précisément des actions de santé publiques et les régressions écologiques peuvent par exemple permettre d'identifier des facteurs de risques méconnus jusqu'alors.

Cependant ces bases de données n'ont pas toujours été conçues pour la recherche épidémiologique plaçant les études dans le contexte du *data reuse*. C'est pourquoi le croisement de ces données se heurte fréquemment au *change of support problem* dans sa composante des *misaligned data zones* du fait de l'absence d'un identifiant spatial unique.

Les solutions au COSP restent peu satisfaisantes car proposent soit de réaliser de larges agrégations dégradant la résolution spatiale et nuisant à la qualité des analyses spatiales soit d'utiliser des modèles de désagrégation spatiale permettant une échelle fine mais entrainant des erreurs d'estimation spatiale des variables. Par ailleurs, il n'a pas été identifié de schéma standardisé de présentation des méthodologies de transformation d'unité spatiales qu'elles soient d'agrégation ou de désagrégation.

Ces deux solutions n'étant pas totalement satisfaisantes, ce travail développe une méthode générique d'interopérabilité des bases de données médicales et écologiques sans utiliser de modèles de désagrégation spatiale tout en conservant l'échelle spatiale la plus fine possible. La méthodologie suivie vise a constituer un cadre standard de présentation d'une démarche de modification d'unité spatiale dans les analyses spatiales en santé.

Puis afin de confirmer la faisabilité de la méthode, des applications sur données réelles médicales et écologiques seront réalisées.

Article

Background

In the field of epidemiology, the term "spatial analysis" refers to the description and analysis of the spatial distribution of healthcare phenomena, such as the incidence or prevalence of disease or healthcare consumption across geographic areas (29,72–75). Although spatial analysis can be applied to point data, geostatistical data and aggregated data, most of the data for spatial analysis in the field of health are aggregated because they ensure that the patients' data remain confidential. By definition, these so-called ecological studies use data that have been aggregated into administrative spatial units, such as counties, provinces and states. These analyses require two categories of aggregated data. The first category is related to how the events (e.g. the cases of disease or surgical acts) are counted within each spatial unit in the study area. The second category is related to the descriptive ecological data on the source population and the living environment within these spatial units, such as the socio-economic level, the employment rate, housing conditions and environmental quality. For example, a spatial analysis of the incidence of Crohn's disease in northern France examined correlations between two data sources: all new cases of Crohn's disease recorded in the EPIMAD register for each district (*canton*), and the characteristics of each of these districts in terms of the underlying population and the living environment. By combining these two sources, the investigators were able to (i) calculate the incidence of Crohn's disease for each *canton*, and (ii) evaluate the influence of the living environment and the population's socio-economic level (33,76).

Spatial analysis in healthcare is attracting growing interest because of improvements in statistical analysis, the development of information technology tools, and the emergence of

disease registries (30,36,40,53,77–79). More recently, the availability of big data in healthcare (52,80,81) and the intensive development of data reuse (82,83) and georeferencing (84,85) have opened up new perspectives for describing healthcare consumption or disease prevalence/incidence over large geographical areas - even whole countries - and analyzing their ecological determinants (such as socio-economic factors) (62,86).

However, the correlation of big data and ecological data over large areas is complicated by the problem of database interoperability (46,51,87). In the specific setting of spatial analysis, interoperability is based on the smallest possible spatial reference unit, which acts as a link between the medical database and the ecological database. In the absence of this link, the data must be aggregated on a larger scale, which limits the precision of the results (54,55,88). In fact, the quality and relevance of the conclusions of a spatial analysis depend on the concordance between the spatial resolution and the nature of the phenomenon studied. The use of aggregated data induces an ecological bias that fades (but does not disappear) when the spatial resolution is increased (9). Moreover, a finer-scale analysis enables the assessment of more local phenomena, such as the impact of sources of pollution (89). However, larger spatial units may be more appropriate if the underlying disease pathways involve larger-scale phenomena. The availability of fine-scale data provides an opportunity to use the scale that best matches the study's goal.

Poor interoperability between medical databases and ecological databases thus appears to be a major limitation for fine-scale spatial analyses of large geographical areas. However, the interoperability problem should not limit the choice of the most appropriate scale. This interoperability problem has been highlighted (for example) for National Health Service data in the UK, Statewide Planning and Research Cooperative System data from New York State

in the USA, and the French national diagnosis-related group database (*Programme Médicalisé des Systèmes d'Information*, PMSI) (48,50,55).

Two ways of tackling the interoperability problem have been suggested: spatial disaggregation and spatial aggregation. The first approach consists in creating a mapping table that adopts the finest scale; consequently, the data aggregated on a larger scale are disaggregated into spatial units at the finest scale. However, this necessitates the use of complex statistical models for spatial disaggregation (such as areal interpolation models) to estimate the variables' values on a smaller scale. Hence, these procedures can lead to errors in the spatial estimation, which are especially large because the spatial units of origin are considered on very different scales (e.g. by going from the state scale to the town scale) (46,90). The second approach (aggregation methods) consists in creating a mapping table that links the spatial units of one or both databases to a larger scale. In a simple, particular case, the data from one of the two databases are aggregated to the spatial scale of the other database. However, in the most frequent case, the spatial units of the two databases are aggregated into a larger spatial unit that covers them both. Although most studies use administrative spatial units as a larger spatial unit, this is not necessarily the finest and/or most appropriate scale for use. Consequently, aggregation methods markedly decrease spatial resolution (e.g. by going from the town scale to the county scale), and may lead to an increase in the ecological bias (54,55,88).

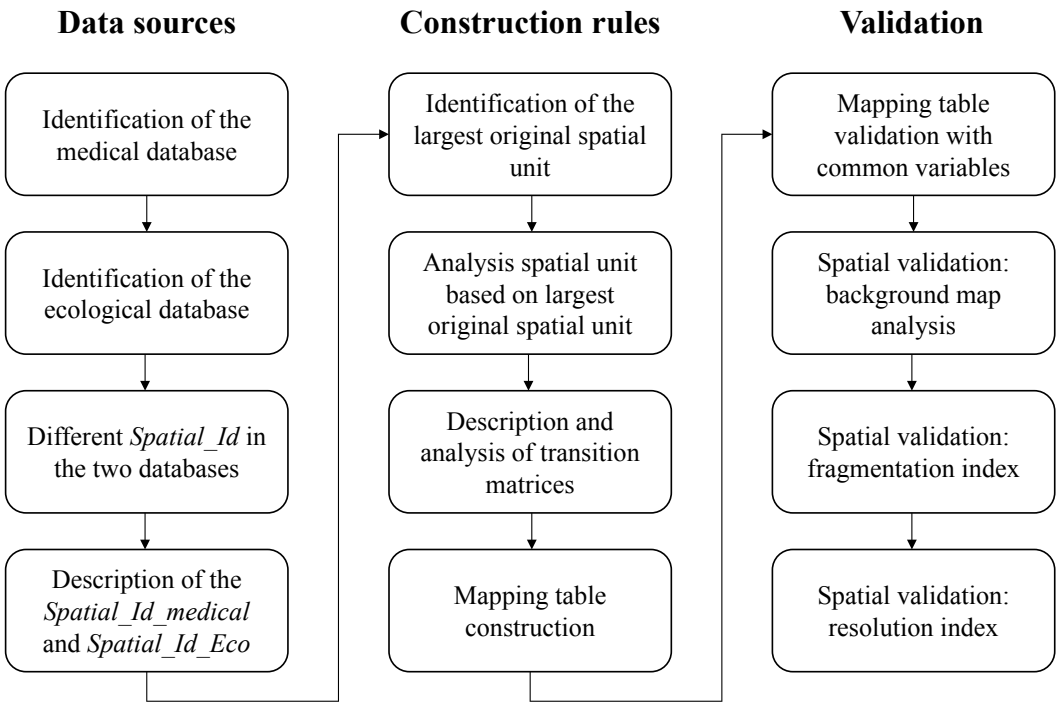
The primary objective of the present study was to develop and characterize a generic method for building a mapping table between a medical database and an ecological database while maximizing the spatial resolution and avoiding the use of spatial disaggregation techniques and thus enabling the choice of most appropriate scale for the phenomenon being studied. By

way of an illustrative example, we applied this method to the interoperability of the above-mentioned PMSI medical database and the socio-economic data produced by the French National Institute of Statistics and Economic Studies (*Institut National de la Statistique et des Études Économiques*, INSEE).

The generic method

This section describes the generic method for improving the spatial interoperability of medical and ecological databases. The different steps in this generic method are summarized in Figure 6.

Figure 6 : A standardized approach for maximizing the interoperability of ecological and medical databases for spatial analysis.



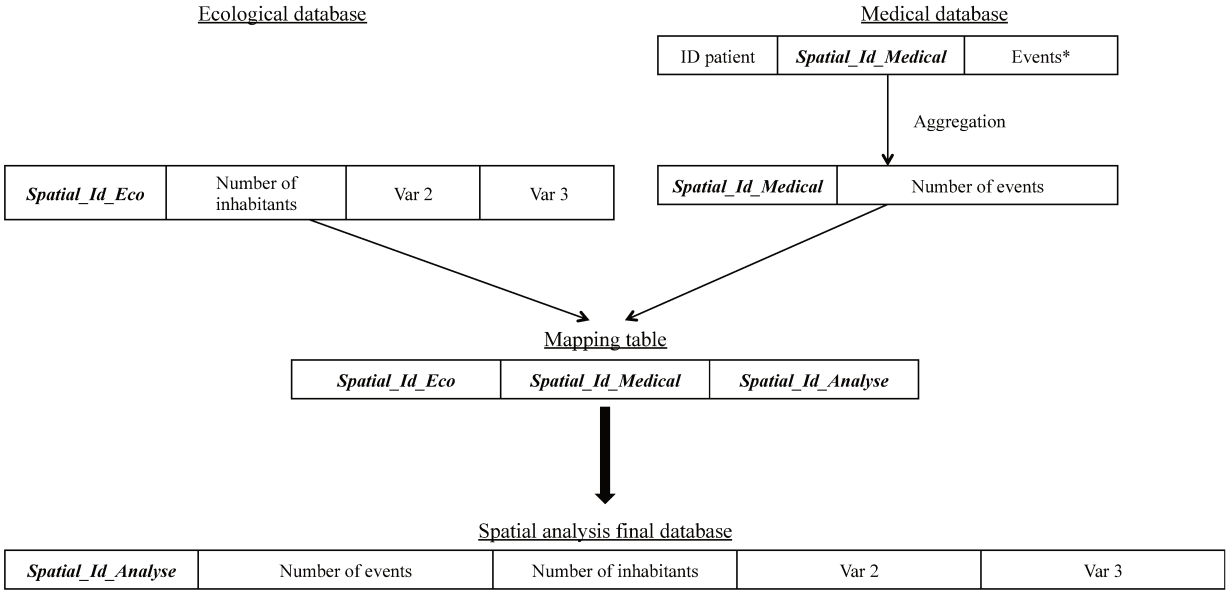
Data and objectives

Let us consider two distinct databases: a medical database that describes patients and healthcare events, and an ecological database that describes the population. The present method considers the following conditions of application:

1. The medical database is organized on the scale of the individual. Each individual is attached to a spatial ID *Spatial_Id_Medical*, which corresponds to the spatial unit *SU_medical*. A variable characterizes each healthcare event.
2. The ecological database is organized on the scale of the spatial unit *SU_eco,* which has a unique spatial ID *Spatial_Id_Eco*.
3. The spatial units *SU_medical* and *SU_eco* differ, as do the spatial IDs *Spatial_Id_Eco* and *Spatial_Id_Medical*.

The objective of our method is to build a mapping table that enables the creation of a final database comprising both medical and ecological data from the above-mentioned databases on the scale of the spatial unit *SU_analysis* and with a unique spatial ID called *Spatial_Id_Analysis*. The medical database must be aggregated for the variable characterizing the healthcare event on the scale of the spatial unit *SU_medical* (Figure 7). An example showing how the final spatial analysis database is built is provided in the Additional file 1.

Figure 7 : A generic method for building a final database for spatial analysis

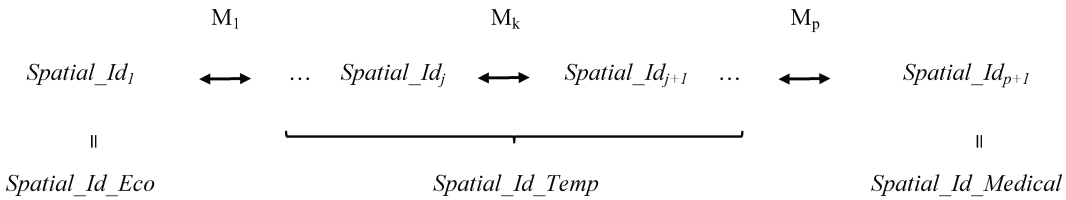


Legend: * a variable characterizing the healthcare event studied (*e.g.:* cases of disease, length of hospital stay, surgical acts, etc.)

Construction rules

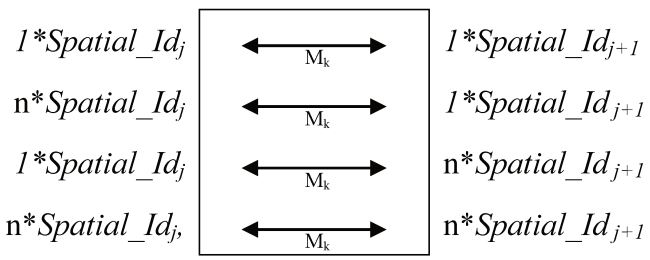
1) ***The direction of the relationship.*** When spatial units differ in size (i.e. *SU_medical* ≠ *SU_eco*), the two databases can only be aligned after the data have been aggregated. Count data are aggregated by calculating a sum, whereas continuous variables or proportions can be aggregated by calculating a median, mean or weighted mean. The larger of the two spatial units is then chosen as *SU_analysis*. The reverse process requires the use of a disaggregation method, leading to a loss of precision (90,91).

2) ***Transition matrices $M_1 \ldots M_p$.*** A transition matrix is a tool for linking an original spatial ID to a final spatial ID:



A mapping table for the IDs *Spatial_Id_Medical* and *Spatial_Id_Eco* IDs can be built by using p transition matrices ($p \geq 1$). For example, a transition matrix makes it possible to associate each town's spatial ID with the spatial ID of the state to which it belongs. However, in more complex situations, there may be no direct way of linking the two spatial IDs. Thus, two or more matrices are required - leading to the creation of at least one temporary spatial ID *Spatial_Id_Temp*. The mapping table yields *p+1 Spatial_Id,* where $Spatial\_Id_1$ corresponds to the *Spatial_Id_Eco* and $Spatial\_Id_{p+1}$ corresponds to the *Spatial_Id_Medical*. The transition matrices are based on a detailed assessment of the *Spatial_Id_Medical* and *Spatial_Id_Eco* IDs. It is then necessary to describe all the equivalence situations for each transition matrix. One or several $Spatial\_Id_j$ can correspond to one or several $Spatial\_Id_{j+1}$ ($1 \leq j < p+1$). The various, mutually exclusive equivalence situations for a given transition matrix M_k ($1 \leq k \leq p$) are shown in Figure 8.

Figure 8 : Mutually exclusive equivalence situations for a transition matrix M_k



Legend: M_k is a transition matrix from among $M_1 \ldots M_p$ transition matrices. M_k describes the relationship between a *Spatial_Id*$_j$ and a *Spatial_Id*$_{j+1}$ and therefore links them. n>1, corresponds to the number of spatial IDs.

Validation

Validation of the mapping table. After the final database has been built, it is necessary to validate the quality of the interface between the medical database and the ecological database. We used the following approach: (i) identification of the set of variables shared by the medical database and the ecological database; (ii) choice of the variables that display the best exhaustiveness and reliability; and (iii) comparison of these variables in the two databases on the scale of the *SU_analysis* spatial unit.

Spatial validation. In spatial terms, the final purpose of the mapping table is to create a background map on the scale of the *SU_analysis* spatial unit. In order to check the quality of the selected spatial unit (*SU_analysis*), it is necessary to evaluate spatial continuity and the decline in spatial resolution.

Spatial continuity is defined as the ability to move from any one point to another point without leaving the spatial unit considered. In other words, a spatially continuous unit has a single boundary (92–94). A spatial unit that does not meet this condition is referred as discontinuous or fragmented. Most studies of putative links between a health outcome and environmental factors rely on the use of aggregated data. These data are frequently represented by the centroid of each spatial unit. However, in the case of discontinuous spatial units, the centroid may be outside the spatial unit. Hence, an error in the data's spatial location (due to fragmented spatial units) might affect the findings and result in an erroneous

conclusion (92–94). In order to control for this eventuality, spatial continuity is evaluated by determining the fragmentation of the spatial units, defined as the number of discontinuous *SU_analysis* as a proportion of the total number of *SU_analysis* (93,94). This index can be calculated using geographical information systems, such as QGIS and ArcGIS (16,17).

Spatial resolution is defined as the surface area of the smallest spatial unit in a given data set; it corresponds to the level of detail within the data. Aggregation of spatial units decreases the spatial resolution and thus the quality of the analysis. For example, the spatial resolution decreases if (for a given geographical zone) the data for a town are aggregated with data for the region as a whole. The decline in spatial resolution can initially be evaluated visually. The background map for *SU_analysis* is compared with the background map for the smallest spatial unit in the initial databases, in order to identify any obviously aberrant aggregates. The decline in spatial resolution can then be measured by calculating the ratio between the median surface area of *SU_analysis* and that of the smallest spatial unit in the initial databases (*SU_initial* = *SU_eco* or *SU_med*). This ratio must also be calculated for other administrative reference units whose surface area is known. These ratios are then compared: a lower index of decline corresponds to a spatial unit with a higher spatial resolution.

$$\frac{\text{SU\_analysis}}{\text{SU\_initial}} \quad \text{VS} \quad \frac{\text{SU\_reference1}}{\text{SU\_initial}} \quad \text{VS} \quad \frac{\text{SU\_reference2}}{\text{SU\_initial}}$$

For example, reference units 1 and 2 could be the county and the state for the USA, or the *canton* and the *département* for France. This index can be also calculated from census data on the number of inhabitants.

Application of the generic method: an illustrative example based on French databases.

Data sources and objectives

In this section, the generic method is applied to a pair of French medical and ecological databases.

1. The medical database is the PMSI. Collection of these data has been approved by the French National Data Protection Commission (*Commission Nationale de l'Informatique et des Libertés*; authorization 1754053). The database is compiled and released by France's Technical Agency for Information on Hospitalization (*Agence Technique de l'Information sur l'Hospitalisation*, ATIH). The database contains a summary of each inpatient stay in France, including the ICD-10 diagnostic code, the medical procedures performed (coded according to the French CCAM classification) and the patient's age, gender, and unique identifier. Each patient is localized by his/her place of residence, which is only characterized by the PMSI spatial ID (*Spatial_Id_PMSI*) in the spatial unit *SU_PMSI*. There were 5644 distinct *SU_PMSI*s in France in 2014, which were characterized by a mean surface area of 97.37 km^2 and a mean population of 11,174.

2. The ecological database was produced by the INSEE (95). The INSEE acts as France's census office, and collects a vast range of demographic, social, economic and housing-related data. Most of the data are publicly available on the INSEE website. The data are summarized for various spatial units: the *commune*, the *canton*, the

département and the *région* (in increasing hierarchical order; see Additional File 2 for details). Most frequently, the data are summarized on the scale of the *commune* (*SU_INSEE*), which is characterized by the spatial ID *Spatial_Id_INSEE*. In 2014, there were 36,594 *communes* (*SU_INSEE*) in France.

3. The spatial units *SU_PMSI* and *SU_INSEE* differ, as do the IDs *Spatial_Id_PMSI* and *Spatial_Id_INSEE*.

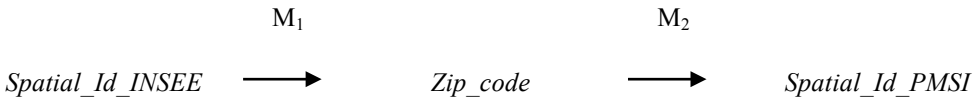
The goal of our method is to create a mapping table for the IDs *Spatial_Id_PMSI* and *Spatial_Id_INSEE*, in order to build a final database that includes both medical data from the PMSI and ecological data from the INSEE. The PMSI medical database provides information on each hospital stay for each patient, which are aggregated for each *Spatial_Id_PMSI* spatial unit. In this illustrative example, the healthcare event of interest is an in-hospital birth. This event was detected by screening for (i) hospital admissions from home, (ii) a patient age of 7 days or less, (iii) admissions from another hospital with a bodyweight below 2500 g, and (iv) admissions from another hospital, with a patient age below 30 days.

Construction rules

The direction of the relationship. The median [interquartile range (IQR)] surface area is larger for *SU_PMSI* (70 km^2 [21.6 – 147.6] than for *SU_INSEE* (10.8 km^2 [6.4 – 18.4]. Accordingly, the spatial unit for the analysis (*SU_analysis*) must be based on the spatial unit *SU_PMSI*, which is characterized by the spatial ID *Spatial_Id_PMSI*.

Transition matrices M_1, M_2.

Two transition matrices were required to establish a correlation between *Spatial_Id_INSEE* and *Spatial_Id_PMSI* via the *Zip_code:*



The various equivalence situations for each transition matrix are presented in the Additional file 3. Transition matrix M1 is obtained by correlating *Spatial_Id_INSEE* (the ID for the *communes*) and the *Zip_code* for the *commune* (96). In France, a zip code corresponds to the geographical zone covered by a single postal delivery office. The equivalence situations are described in detail in Table 4. In over 95% of cases, a given zip code covers several *communes*, which leads to the first data aggregation step (Table 1: situations 1, 4 and 5). In large, highly populated *communes* (<1%), many zip codes correspond to a single commune. Each zip code corresponds to a single subset of the *commune*, and the union of these distinct subsets constitutes a *commune* (situation 2). In 5% of cases, the zip code corresponds to the *commune*'s *Spatial_Id_INSEE* (situation 3).

Transition matrix M2 is obtained by correlating *Spatial_Id_PMSI* and *Zip_code*. According to the ATIH, *Spatial_Id_PMSI* has to be built from zip codes for legal reasons (97). Thus, *Spatial_Id_PMSI* is equivalent to the zip code's geographic area when the level of statistical confidentiality is high enough (in over 99% of cases; situations 1, 3 and 5). In the opposite case, *Spatial_Id_PMSI* corresponds to the aggregation of several zip codes (<1% of cases: situations 2 and 4). A second aggregation step is then performed. In situation 2, the transition matrix M1 connects the *commune* to several zip codes. However, data partition is not necessary because the transition matrix M2 aggregates exactly the same units.

Lastly, a *Spatial_Id_analysis* ID is attributed to each of the *Spatial_Id_PMSI* (situations 1 to 4). The combination of transition matrix M1 and transition matrix M2 can, however, generate

a small number of particular cases (<1% of cases). In situation 5, several *Spatial_Id_PMSI* IDs have at least one *Spatial_Id_INSEE* ID in common. It is then impossible to obtain an exact correlation between the spatial ID from the PMSI and the spatial ID from the INSEE. In this situation, the *Spatial_Id_PMSI* IDs are aggregated into a single *Spatial_Id_Analysis* ID. Thus, 23 *Spatial_Id_PMSI* IDs were grouped into 11 *Spatial_Id_Analysis* IDs. In total, there were 5632 *Spatial_Id_Analysis* IDs in the final database.

The data processing and statistical analyses were performed using R software (version 3.3.2) (18). QGIS software (version 2.14) was used to create the background map and calculate the fragmentation index (17).

Table 4 : Mapping table for *Spatial_Id_INSEE* and *Spatial_Id_PMSI*.

Situations	Mapping table						Proportion (%)	
	Spatial_Id_Eco	M_1	*Spatial_Id_Temp*	M_2	*Spatial_Id_Medical*	*Spatial_Id_Analysis*	*Spatial_Id_INSEE* (n = 36594)	Number of inhabitants (n = 63375971)
1	*Spatial_Id_INSEE* *Spatial_Id_INSEE* *Spatial_Id_INSEE*	→	*Zip_code*	→	*Spatial_Id_PMSI*	→ *Spatial_Id_Analysis*	94.5 (n=34602)	56 (n= 35260461)
2	*Spatial_Id_INSEE*	→ → →	*Zip_code* *Zip_code* *Zip_code*	→ → →	*Spatial_Id_PMSI*	→ *Spatial_Id_Analysis*	<1 (n= 48)	9 (n= 5707929)
3	*Spatial_Id_INSEE*	→	*Zip_code*	→	*Spatial_Id_PMSI*	→ *Spatial_Id_Analysis*	4.7 (n=1719)	35 (n= 22016678)
4	*Spatial_Id_INSEE* *Spatial_Id_INSEE* *Spatial_Id_INSEE* *Spatial_Id_INSEE*	→ → → →	*Zip_code* *Zip_code*	→ →	*Spatial_Id_PMSI*	→ *Spatial_Id_Analysis*	<1 (n=208)	<1 (n= 98684)
5	*Spatial_Id_INSEE* *Spatial_Id_INSEE* *Spatial_Id_INSEE* *Spatial_Id_INSEE*	→ → → →	*Zip_code* *Zip_code* *Zip_code*	→ → →	*Spatial_Id_PMSI* *Spatial_Id_PMSI* *Spatial_Id_PMSI*	→ *Spatial_Id_Analysis*	<1 (n=14)	<1 (n= 292219)

Validation

Validation of the mapping table. In order to evaluate the quality of the match between the PMSI database and the INSEE database, the annual number of live births was used as the common variable.

The number of births associated with each *Spatial_Id_INSEE* ID was provided by the INSEE. The number of births associated with each *Spatial_Id_PMSI* ID was obtained by extracting the PMSI database.

For each *Spatial_Id_Analysis* ID, the indicators were compared by calculating the relative difference (i.e. the difference between the number of births in the INSEE data and the number of births in the PMSI data, divided by the number of births in the PMSI data). These relative differences are quoted as the median [IQR]. The total number of births was 785,742 in the INSEE database and 737,545 in the PMSI database, giving a difference of 48,197. The median [IQR] relative difference was 2.3% [0 – 5.7] (a boxplot is available in the Additional file 4).

In 2012, the ATIH performed an extensive study of the number of inhabitants in each *SU_PMSI* spatial unit, based on the INSEE data. The data on the number of inhabitants are available online for each *Spatial_Id_PMSI* (98). We therefore transformed these data on the scale of the *Spatial_Id_Analysis* and compared the population data provided by the ATIH and the population data provided for *SU_INSEE*, as aggregated by our mapping table. For each of the *Spatial_Id_Analysis* IDs, the correlation was perfect (difference = 0). Hence, the resulting mapping table automatically performs the task described by the ATIH, regardless of the INSEE variable.

Spatial validation.

A background map of the *SU_analysis* spatial unit was created using data from the French National Geographic Institute (*Institut National de l'Information Géographique et Forestière*, IGN) (Figure 9). Spatial continuity was evaluated by calculating the fragmentation index; this was 2.4% (n=134) for the 5632 *SU_analysis* spatial units. This value is within the range of fragmentation indices (2% to 40%) reported for public use microdata areas (PUMAs) in the USA [37].

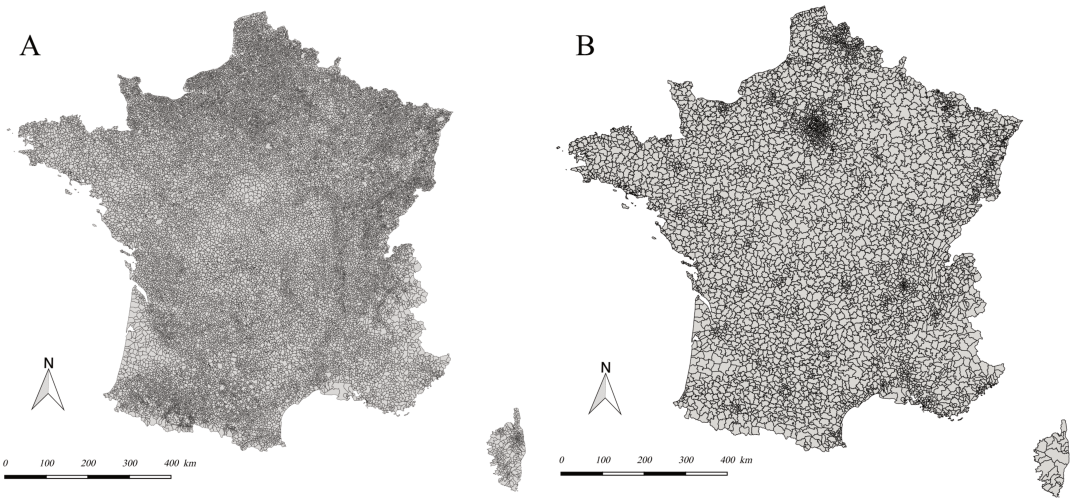
A possible decline in the spatial resolution was evaluated first by visual comparison of the respective background maps for *SU_INSEE* and *SU_analysis* (Figure 9). The *SU_analysis* spatial unit appeared to be regularly distributed over the geographical zone, with no aberrant aggregations. In a second step, we calculated the decline index for the spatial resolution required to obtain data on the surface area and the number of inhabitants in the French *communes*, *cantons* and *départements*. The surface area data came from the IGN, whereas the data on the number of inhabitants came from the INSEE database. The comparison of the spatial resolution index for *SU_analysis* with the French administrative units is described in Table 5. The *SU_analysis* unit has a lower decline index than the *cantons* and *départements* for the surface area (6.5, 13.6 and 555.3, respectively) and the number of inhabitants (14.6, 24 and 1249.2, respectively).

Table 5: Comparison of the numbers of inhabitants and surface areas.

	N	Surface area[a]			Number of inhabitants[b]		
		Index[c]	Median	IQR[d]	Index[c]	Median	IQR[d]
SU_reference							
Communes	**36594**	**1**	**10.8**	**6.4 – 18.4**	**1**	**0.4**	**0.2 – 1.1**
Cantons	**3708**	**13.6**	**146.2**	**66 – 209.6**	**24**	**10.4**	**5.5 – 20.4**
Départements	**96**	**555.3**	**5986**	**5153 – 6811**	**1249.2**	**540.9**	**306.5 – 855.8**
SU_analysis	**5632**	**6.5**	**70**	**21.6 – 147.6**	**14.6**	**6.3**	**3.5 – 11.8**

Legend: Comparison of the numbers of inhabitants and surface areas for French administrative spatial units and the *SU_analysis* spatial unit, via calculation of the decline index for spatial resolution (2014 data). a: Surface area (in square kilometres); b: Number of inhabitants (in thousands); c: Ratio between the median for the spatial unit and the median for the *commune*; d: Interquartile range.

Figure 9 : Background map for the *SU_INSEE* (A) and *SU_analysis* (B) in mainland France in 2014.



Legend: (A): Background map of the spatial unit *SU_INSEE*, which represents the French *communes* in 2014. (B): Background map of the spatial unit *SU_analysis*, which represents the analysis spatial unit for our application in 2014.

On average, the spatial unit for analysis is therefore 6 times larger than the smallest available unit, testifying to a loss of spatial resolution. However, our method minimizes this loss; for the surface area, the scale is twice as fine as the first reference unit (the *canton*) and nearly 100 times finer than for the *département* (the second reference unit).

As an illustrative example of the application of this method, the birth rate was mapped (Additional file 5).

Discussion

The method presented here addresses the interoperability problem for ecological and medical databases in a context of the spatial analysis of healthcare events. The loss of spatial resolution was minimized, and we did not have to resort to the use of spatial disaggregation techniques. The method's application to French national data enabled us to correlate medical data from the PMSI database with ecological data from the INSEE database - resulting in the creation of a final database for fine-scale spatial analysis.

This method may be of value for correlating ecological and medical big data in spatial analyses. This type of data is increasingly available and is opening up new perspectives in epidemiology. However, the use of medical big data in the field of spatial analysis is restrained by interoperability problems, known as the change-of-support problem and the misaligned data problem (46). In Rossheim et al.'s study of alcohol sales and the socio-economic environment, the data were available on the scale of the zip code, the census block or the zip code tabulation area. To perform analyses on the zip code scale, the researchers were obliged to use spatial disaggregation and aggregation methods; this decreased the quality

of the final spatial analysis database (99). A similar problem was encountered in Sundmacher and Busse's study of the link between physician supply and avoidable cancer deaths in Germany. The lack of interoperability and the broad range of ecological databases prompted the researchers to use spatial interpolation methods on the district level and to not integrate certain environmental data – thus placing limitations on their analyses (100).

Spatial resolution is a major issue in the spatial analysis of healthcare data because it is easier to detect local phenomena when the resolution is high (9). For example, the decrease in spatial resolution affects the precision with which a cluster can be localized (53,101). The variation in the results of a spatial analysis as a function of the spatial resolution was emphasized by Lee et al.'s study of obesity in the USA; fewer healthcare events were identified when the spatial resolution fell (102). Jeffery et al. came to a similar conclusion in their study of paediatric leukemia (103).

The advantage of our method consists in opting for aggregation on the finest scale possible, whilst checking the quality of the final spatial analysis database. This approach appears to have been used previously in a study of stroke, although the method's details were not specified (62). The use of spatial disaggregation methods is not desirable, since they lead to a loss of precision in spatial analysis - even when complex models are used (46,90). Furthermore, validation of the mapping table results in a high-quality final database for spatial analysis. The spatial validation process ensures that the greatest possible spatial resolution is achieved. Lastly, validation ensured that the spatial units' fragmentation index remains low. By way of an example, Siordia et al.'s studies of the American PUMA database featured a high fragmentation index and thus encountered theoretical difficulties in the application of statistical models; the spatial position of a healthcare event was no longer coherent with that

of a spatial unit (93,94). This generic method may provide a structural framework so that researchers can provide a standardized description of the methods used to aggregate ecological and medical spatial data.

Nevertheless, our present method has a number of limitations, most of which are inherent to all spatial analyses. Firstly, a large percentage of the scenario 5 (Table 4) might decrease the spatial resolution, due to the aggregation of several basic spatial units. This issue can be evaluated by analyzing the spatial resolution index (as presented in the present study) and establishing whether the final spatial unit sizes are homogeneously distributed or not. Secondly, the geographical boundaries of spatial units change over time, which can make it more difficult to study healthcare events over a long time interval. This problem can be tackled in two ways: by optimizing the study period and thus minimizing changes in geographical boundaries or by considering the geographic boundaries that correspond to the longest study period. Thirdly, our method only partly addressed the change-of-support problem because it only applies to aggregated data (a frequent situation in the spatial analysis of healthcare events, nevertheless) (104). Therefore, for other types of spatial data (such as geostatistical data), preliminary work on aggregation to the spatial unit of interest must be carried out in collaboration with specialists in the particular field. Lastly, the present method requires the definition of transition matrices prior to construction of the mapping table.

Conclusion

In conclusion, the present work suggests that it is possible to significantly improve the interoperability of ecological databases and medical databases, and thus enable finer-scale analyses. In view of the growing availability of big data, the method presented here could be a useful tool for the precise spatial analysis of large geographical areas.

Résultats complémentaires

Cette méthode générique permet de croiser les bases de données du PMSI avec tout type de bases de données à l'échelle communale. Par ailleurs, elle fournit un cadre d'étude systématique pour les situations de COSP avec une démarche en 3 étapes : analyse des sources de données, établissement des règles de construction de la table de correspondance et enfin validation. Ainsi, plusieurs pistes d'applications de cette méthode ont pu être testées au niveau national : un calcul du taux de natalité nécessitant les données de naissance du PMSI et les données de populations de l'Insee, une mesure d'un indice de défaveur sociale, le French EDI, à l'échelle du *Spatial_Id_PMSI* et enfin une étude épidémiologique préliminaire sur la répartition spatiale de l'incidence de la fracture de hanche en France.

I. Taux de natalité

A. Introduction

Le taux de natalité a été utilisé comme piste d'application car les données sont facilement accessibles que ce soit dans les bases de l'Insee ou dans les bases du PMSI. Par ailleurs, les données de naissances font l'objet d'une littérature de plus en plus importante portant sur leurs liens avec les facteurs socio-économiques et environnementaux. L'étude de Coker et al montrant un lien entre faible poids de naissance et microparticule de moins de 2,5nm en est un exemple (25). C'est pourquoi le calcul d'un taux de natalité était un moyen efficace et efficient de montrer qu'il était possible de croiser les données de naissance du PMSI avec une donnée écologique : la population.

B. Méthodes

i. Sources de données

Les données du PMSI provenaient d'une extraction sur toute la France pour l'année 2012. Les codes Z38* étaient utilisés comme identifiant d'une naissance. Ils sont en effet codés pour tous les séjours de naissances comme indicateur de périnatalité (49) et leur qualité est reconnue (105). L'identifiant spatial du PMSI était le code géographique PMSI. Les séjours du PMSI avec des codes géographiques non pris en compte (étranger, Dom-Tom) ou erronés étaient exclus de l'analyse. Les enfants né sans vie identifiées par le groupe homogène de malade « 15Z10E » étaient exclus de l'analyse. Les séjours correspondant à un même patient dans la base ont été supprimés à l'aide de l'identifiant anonyme « id_ano ».

Les données de l'Insee provenaient des bases démographiques en accès libre. Elles correspondaient au nombre de naissance vivantes et à la population par commune en France métropolitaine pour l'année 2012. L'identifiant spatial de chaque commune était le code géographique INSEE.

ii. Construction de la table de correspondance

La situation était celle rencontré lors du développement de la méthode générique. L'unité spatiale PMSI est plus grande que l'unité spatiale de la commune et a été utilisée pour construire les identifiants spatiaux finaux *Spatial_Id_analyse.*

iii. Validation

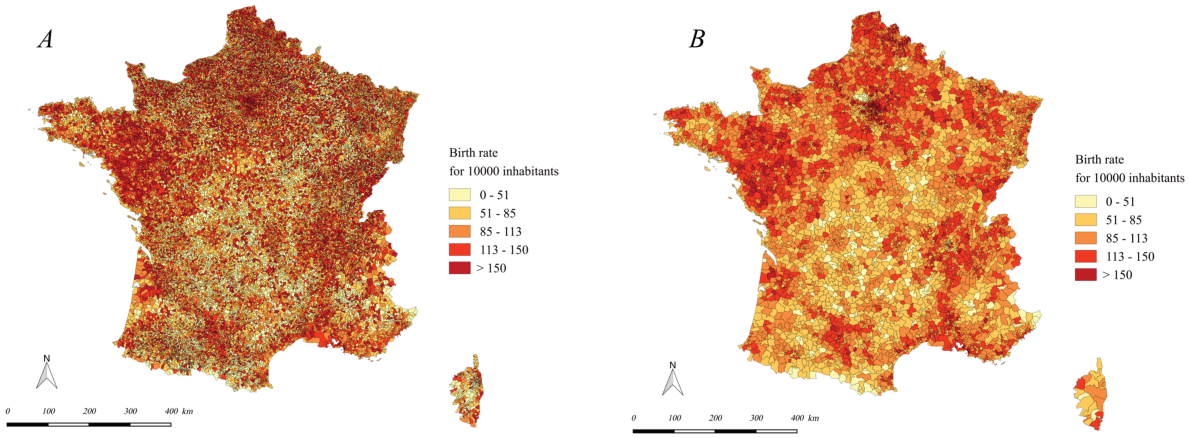
La validation de la table de correspondance et la validation spatiale ont été effectuées dans le développement de la méthode générique.

Dans l'objectif d'une validation de l'analyse, la carte finale des naissances était comparée à celle des taux de natalité communaux obtenus à partir des données de l'Insee.

C. Résultats

La Figure 10 présente le taux de natalité comparé à l'échelle communale et à l'échelle des *SU_analyse.* Les échelles des cartes sont harmonisées et les données sont visuellement comparables. Le taux de natalité est dans les deux situation plus important dans le Nord et l'Est de la France. Les deux cartes sont concordantes l'une avec l'autre.

Figure 10 : Taux de natalité pour 10000 habitants en France métropolitaine calculé à partir des données de l'Insee à l'échelle communale (A) et des données du PMSI à l'échelle des *SU_analyse* (B).



D. Discussion

Il a été possible dans cette application de croiser des données médicales de naissances du PMSI avec des données écologiques de population de l'INSEE et d'en réaliser une représentation cartographique à l'échelle des *Spatial_Id_analyse*.

Il y avait une bonne concordance entre les deux cartes ce qui suggérait une bonne efficacité de la table de correspondance pour traiter ces données et croiser données du PMSI et données de l'INSEE.

Cependant il peut exister des différences dues au fait que les sources de données ne sont pas identiques. Le PMSI ne contient par exemple que les naissances ayant fait l'objet d'une hospitalisation. Par ailleurs, le PMSI peut contenir des erreurs tels que l'absence de codage du code Z38 pour une naissance. Enfin les séjours au code géographique PMSI erroné ont été supprimés.

En conclusion, malgré les limites de l'utilisation du PMSI pour l'analyse des naissances, il a été possible d'appliquer la méthode générique permettant le croisement de données médicales et écologiques en conservant une échelle fine.

II. French EDI

A. Introduction

Pornet et al. ont développé un indicateur composite européen adaptable à chaque pays permettant d'évaluer la défaveur sociale (5). Cet indicateur est aujourd'hui largement utilisé et étendue à plusieurs pays européens (106). Le calcul du French EDI utilise des items tels que le taux de chômage, l'équipement en voiture, le niveau d'étude ou la taille de la famille. Plus le French EDI est élevé pour une unité spatiale donnée, plus la zone correspondante est considérée comme défavorisée. La défaveur sociale est un facteur de risque important de nombreuses pathologies (107). Il était utile de savoir s'il était possible de l'utiliser à l'échelle du code géographique PMSI.

B. Méthodes

i. Sources de données

Les données sont celle de la base de données fournie par l'équipe de Pornet et al. Cette base de données contient l'ensemble des French EDI par commune pour l'année 2011.

ii. Construction de la table de correspondance

La table de correspondance présentée dans l'étude de développement de la méthode générique a été utilisée pour agréger les données à l'échelle de la commune en données à l'échelle du *Spatial_Id_analyse*, construit à partir du code géographique PMSI. Le calcul d'une médiane a été utilisé pour réaliser l'agrégation.

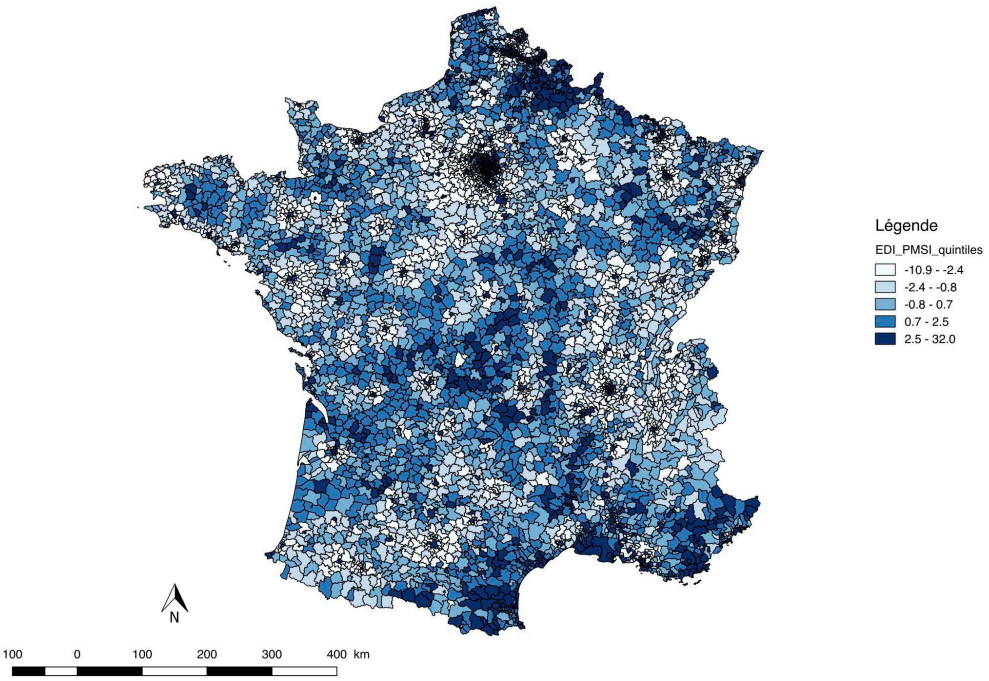
iii. Validation

La validation de la table de correspondance ainsi que la validation spatiale des *Spatial_Id_analyse* ont été réalisées lors du développement de la méthode générique.

C. Résultats

Les données communales ont été transformées grâce à la table de correspondance en données à l'échelle *SU_analyse*. La Figure 11 présente le French EDI pour chaque *Spatial_Id_analyse* de France métropolitaine hors Corse.

Figure 11 : French EDI pour chaque *Spatial_Id_analyse* en 2011 en France métropolitaine hors Corse.



D. Discussion

Le French EDI ayant pu être calculé pour chaque *SU_analyse*, il est possible de l'utiliser dans des analyses plus complexes telles que des régressions écologiques afin de chercher des liens entre défaveur sociale et survenue de pathologies. Cependant cet indice de défaveur sociale n'est pas le seul qui existe et ne permet pas de rendre compte correctement de tous les éléments et en particulier des différences entre le tissu urbain et les milieux ruraux comme l'a soulignée Pornet et al. (5).

III. Fracture de l'extrémité supérieure du fémur.

La fracture de l'extrémité supérieure du fémur (FESF) est la première application épidémiologique de la méthode générique. Tout d'abord les bases scientifiques de l'étude seront présentées puis la méthode utilisée sera détaillée. Enfin les résultats préliminaires connus à ce jour seront présentés ainsi que les nuances qu'il est possible d'y apporter.

A. Introduction

i. Intérêt de ce sujet d'étude

La FESF représente un problème de santé publique important pour les personnes âgées. En effet, cette population en présente fréquemment d'une part du fait d'un risque de chute important et d'autre part du fait d'une fragilité osseuse s'aggravant avec l'âge (108,109). Le traitement de cette pathologie passe le plus souvent par une chirurgie lourde d'ostéosynthèse ou de mise en place d'une prothèse totale de hanche. Cette chirurgie entraine une morbidité importante avec un fort risque de grabatisation ainsi qu'un risque de décès non négligeable (110–112). Par ailleurs, la mise en place de matériel, le suivi ainsi que les hospitalisations et ré-hospitalisations ont un coût important pour la société bien que cette stratégie semble être « coût-efficace » (113). Les fractures ostéoporotiques de hanche sont connues comme étant

liées à la situation socio-économique des populations (114). Cette situation socio-économique peut être estimée par l'indice de défaveur sociale qu'est le French EDI. Cependant aucune étude à ce jour n'a pu être réalisée sur l'ensemble du territoire en France. C'est pourquoi l'analyse spatiale sur les données nationales du PMSI est un bon outil pour connaître la répartition de la pathologie en France et en mesurer le lien avec la défaveur sociale.

ii. Objectif

Le premier objectif de cette étude est d'étudier la répartition spatiale de l'incidence FESF chez la personne âgée au niveau national. Le second objectif est d'étudier l'association entre cette répartition spatiale et un indice de défaveur sociale (French EDI).

B. Méthodes

i. Sources de données

La base de données des patients provenait d'une extraction de la base PMSI nationale de 2007 à 2014. L'analyse portait sur les séjours de 2012 à 2014 de patients de plus de 50 ans résidant en France métropolitaine présentant les codes diagnostics PMSI appartenant à la liste des codes CIM-10 disponibles en Annexe 6.

Les séjours présentant les codes diagnostics CIM-10 fournis en Annexe 7 ou ayant eu des antécédents sur la période 2007-2011 étaient exclus de l'analyse. Ces éléments visaient à ne conserver uniquement les patients ayant eu une fracture de hanche traumatique.

ii. Construction de la table de correspondance

La table de correspondance utilisée dans le développement de la méthode générique a été réutilisée. Cependant quelques adaptations ont été nécessaires. En effet les données du PMSI étaient pluri-annuelles et certains codes géographiques PMSI utilisés n'étaient pas pris en compte dans la table de correspondance dans les cas de zones peu densément peuplées. Les

corrections nécessaires ont été apportées à la table de correspondance. Les *Spatial_Id_analyse* étaient rigoureusement les mêmes que dans le développement de la méthode générique.

iii. Validation

Les corrections n'ayant pas impacté les unités spatiales d'analyse, les validations effectuées dans le développement de la méthode générique étaient correctes.

iv. Plan d'analyse

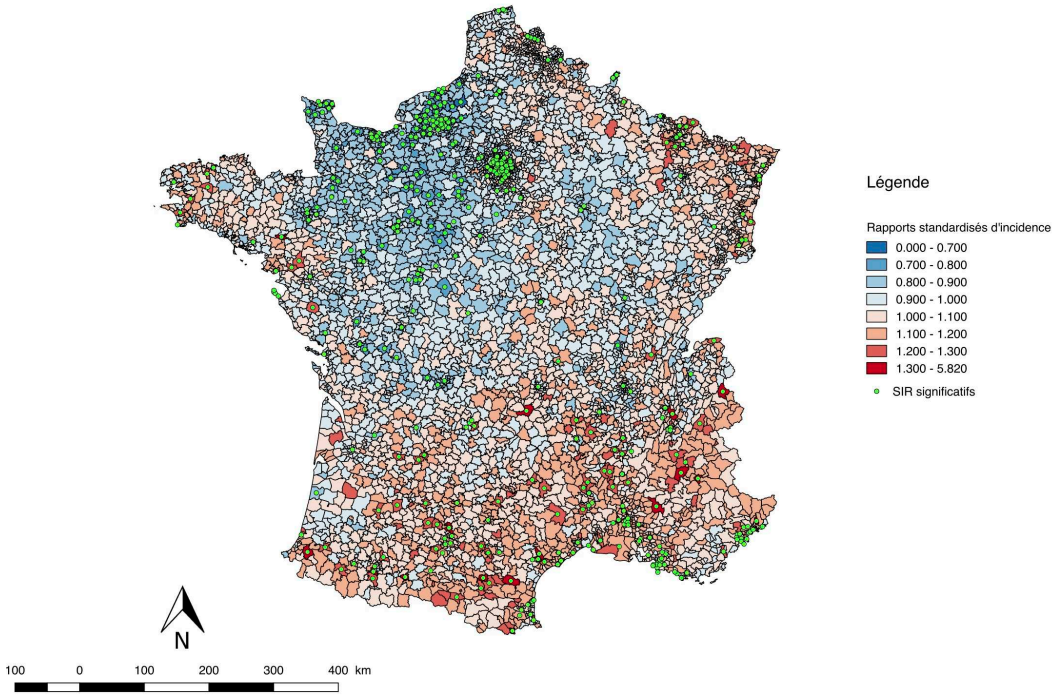
Trois méthodes d'analyse spatiale sont proposées ici. Tout d'abord les SIR lissés par méthode bayésienne ont été calculés et permettent d'estimer un risque relatif par unité spatiale. L'âge et le sexe sont des facteurs de confusion potentiels (115). C'est pourquoi les SIR ont été ajustés sur l'âge et le sexe par standardisation indirecte basée sur les données de populations de l'INSEE à l'échelle communale en 2013. Ensuite, les statistiques de scan spatiales de Kuldorff ont été utilisées afin de mettre en évidence des zones de sur- ou de sous-incidence (36,37). Enfin, une régression écologique basée sur le modèle BYM a été utilisée afin de déterminer les facteurs prédictifs de FESF en associant en particulier un risque relatif au French EDI. Cette variable initialement disponible sous forme quantitative continue a été transformée en variable qualitative à 5 modalités (quintiles).

C. Résultats

La table de correspondance a permis de transformer les données de population de l'Insee à l'échelle de la commune en données disponibles à l'échelle du *Spatial_Id_analyse.* Il a été possible de croiser ces données avec les données du PMSI. La représentation cartographique des risques relatifs estimés par calcul des SIR lissés par méthode bayésienne est présentée en Figure 12. Les unités spatiales présentant une sur-incidence de FESF (SIR>1) apparaissent en rouge alors que les unités spatiales présentant une sous-incidence de FESF (SIR<1) apparaissent en bleu. Les points verts correspondent aux SIR significatifs, c'est à dire dont

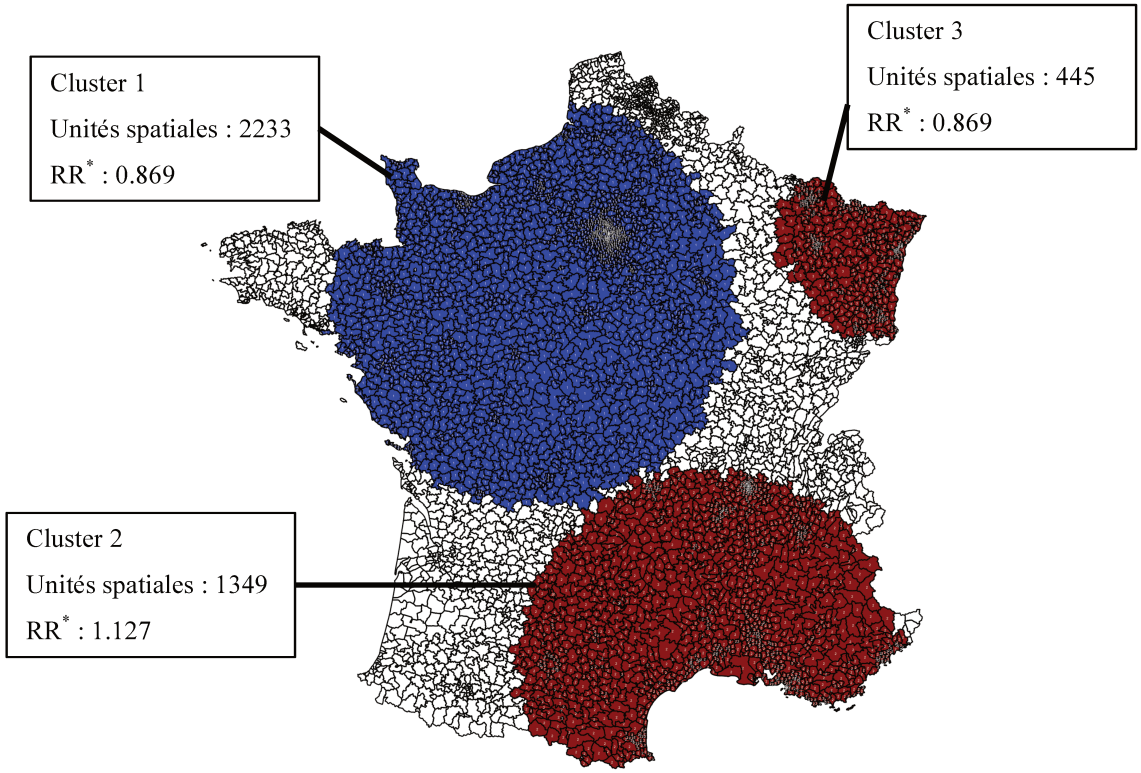
l'intervalle de crédibilité ne contient pas la valeur 1 (risque neutre). Il semble exister un gradient Nord-Sud avec des zones de sur-incidence essentiellement localisées dans le Sud de la France.

Figure 12 : Rapports standardisés d'incidence lissés par méthode bayésienne ajustés sur l'âge et le sexe de fracture de l'extrémité supérieure du fémur en France métropolitaine hors Corse. Base nationale du PMSI. 2012-2014



Les statistiques de scan spatiales ont permis de mettre en évidence un cluster de sous-incidence et deux cluster de sur-incidence présentés en Figure 13. Le cluster de sous incidence se situe dans le nord de la France et présente un risque relatif de 0.869. Le plus grand cluster de sur incidence se trouve dans le sud de la France et présente un risque relatif de 1.127. Ces résultats sont ainsi comparables aux résultats des SIR lissés.

Figure 13 : Clusters spatiaux de sur-incidence (rouge) et de sous-incidence (bleu) de FESF en France métropolitaine hors Corse de 2012 à 2014.



* : RR : Risque relatif

Enfin le calcul de la régression écologique a permis de chercher un lien entre le French EDI analysé en quintile et la survenue de FESF. Ces résultats sont présentés dans le Tableau 6. Le 1er quintile utilisé comme référence représente la valeur la plus faible du French EDI. Les 3ème, 4ème et 5ème quintile sont significativement associés à une augmentation du risque de survenue de FESF.

Tableau 6 : Résultat de la régression écologique entre le French EDI catégorisé en quintiles et la survenue de FESF en France métropolitaine.

Quintiles du French EDI	Risques relatifs	Intervalle de crédibilité à 95%
alpha	0.954	[0.938 ; 0.970]
2ème quintile	1.006	[0.986 ; 1.026]
3ème quintile	1.030	[1.008 ; 1.051]
4ème quintile	1.028	[1.007 ; 1.050]
5ème quintile	1.021	[1.0001 ; 1.042]

D. Discussion

Cette application a permis de démontrer qu'il était possible d'utiliser de manière aisée la méthode générique ainsi que la table de correspondance pour analyser les données du PMSI. Il a ainsi été possible de réaliser la cartographie de la fracture supérieure de l'extrémité du fémur ainsi que de chercher un lien avec le French EDI. La cartographie réalisée est d'ailleurs comparable avec le résultat présenté dans une étude de la DREES qui retrouvait les mêmes zones de sur- et sous-incidence à l'échelle départementale (54).

Si ce lien a bien été identifié, il reste ténu. Cela peut provenir d'un décalage entre les données concernant spécifiquement la population des personnes âgées de plus de 50 ans et le French EDI calculé en population générale. Il est tout à fait possible que le niveau de défaveur sociale de la population générale d'une unité spatiale soit différent du niveau de défaveur sociale réelle de la population âgée. C'est particulièrement possible dans la société occidentale dans laquelle les personnes âgées ne vivent pas avec la famille et tendent à être isolées (116,117). Il serait intéressant d'adapter le French EDI à la personne âgée ou de créer un nouvel indicateur

de défaveur sociale spécifique à la personne âgée afin d'être en mesure de correctement prendre en compte ces paramètres dans de futures analyses.

Il s'agissait ici d'un exemple d'application qui visait à tester la faisabilité et la rapidité d'exécution de la méthode. Les questions cliniques et épidémiologiques sur la FESF restent entières et doivent mener à la réalisation de nouvelles analyses.

Discussion générale

I. Résultats principaux

La méthode développée dans ce travail a tout d'abord permis de construire une table de correspondance entre une base de données écologiques et une base de données médicales. Ensuite, cette méthode peut être répliquée selon un schéma standard sur d'autres bases de données et permet de ne pas utiliser de méthode de désagrégation spatiale tout en permettant de conserver la plus fine échelle possible. Enfin, les applications sur différentes données ont montré que ce travail méthodologique peut aisément être utilisé en pratique courante.

II. Un cadre d'utilisation pour le traitement des grandes bases de données spatiales

Cette utilisation s'inscrit dans la boîte à outils de l'analyse spatiale en santé. En effet, il existait un manque dans la méthodologie de traitement et de transformation des données pour l'analyse spatiale. Or, l'objectif de reproductibilité de la recherche dans le contexte de l'*Evidence Based Medicine* nécessite que la méthodologie soit la plus expliquée possible afin que les pairs puissent réaliser un travail similaire avec les mêmes bases de données. L'étude de Roussot et al. analysait par exemple des données sur les accidents vasculaires cérébraux en France sans expliquer les méthodes utilisées pour la transformation des données (62). Il était impossible de répliquer les résultats de cette analyse et d'en vérifier la validité.

Cette problématique est particulièrement forte dans le cas des grandes bases de données qui sont de plus en plus fréquentes. En effet, le développement du géoréférencement ainsi que la dynamique de l'ouverture des données rendent la mise à disposition de grandes bases de données de plus en plus simple. Le traitement informatique d'informations géographiques est facilité par les outils actuels tels que le logiciel R qui permet, entre autres, d'automatiser le

géocodage d'adresses à partir des données de la poste mise à disposition à travers la base d'adresse nationale ou de Google Maps (118). Des projets tels qu'*OpenStreet Map* mettent aussi à disposition un très grand nombre de données géographiques telles que le réseau routier ou les limites administratives (119). Il existe aussi des données environnementales françaises sur l'ensemble des sites polluants ou potentiellement polluants à travers les bases de données Basias et Basol. Le croisement de ces bases de données avec les données médicales peut amener à l'identification de facteurs de risques de pathologies variées.

La méthode proposée dans ce travail propose un cadre d'étude permettant d'assurer l'interopérabilité de ces grandes bases de données qu'il est capital de manipuler et d'interconnecter efficacement.

III. Choisir la résolution spatiale la plus adaptée

En introduction de ce travail, il a été expliqué comment la résolution spatiale influe sur les résultats de l'étude et leur interprétation. Le biais écologique est impacté par les modification de résolution spatiale. De même, la performance des analyses pour mettre en évidence des phénomènes locaux est sensible à ce paramètre. La littérature disponible sur les statistiques de scan spatiales est un riche exemple des liens entre résultats de l'analyse et résolution spatiale (53,120).

Le choix de la résolution spatiale suit deux critères :

- L'objectif scientifique de l'étude, c'est à dire l'adéquation entre la résolution spatiale utilisée et la nature du phénomène étudié ;
- Les données disponibles, c'est à dire les différentes résolutions spatiales accessibles au chercheur.

La situation idéale se présente lorsque seul l'objectif scientifique est pris en compte. Ce dernier peut parfois nécessiter une résolution très fine mais parfois une résolution spatiale plus large est nécessaire en fonction de l'hypothèse de recherche. La méthode générique proposée dans ce travail permet d'utiliser si nécessaire la résolution spatiale la plus fine possible sans avoir recours à des méthodes de désagrégation spatiale. En l'utilisant, le chercheur dispose d'une unité spatiale à fine échelle qu'il a possibilité d'agréger selon la résolution spatiale la plus adaptée à l'objectif de l'étude. Ainsi, le chercheur dispose d'un plus grand choix de résolutions spatiales et peut privilégier l'objectif scientifique de l'étude.

De plus, la méthode propose de se passer des méthodes de désagrégation spatiale. Ces dernières permettent de transformer des données d'une échelle A en une échelle B plus fine. Elles ont cependant pour inconvénient d'entrainer des erreurs d'estimations des données à l'unité spatiale d'analyse.

Enfin, les critères de validation spatiale permettent de confirmer les qualités intrinsèques de l'unité spatiale finale. La fragmentation des unités spatiales peut induire des imprécisions dans les analyses spatiales que cela soit dans la définition des voisins ou le calcul du centroid d'une unité spatiale. C'est pourquoi il faut limiter cette fragmentation et contrôler ce paramètre. Par ailleurs, la dégradation de la résolution spatiale doit être mesurée car l'un des objectifs initiaux de la méthode est de limiter au maximum l'agrégation. *In fine,* l'unité spatiale est validée ce qui permet d'appuyer les résultats des analyses spatiales.

IV. Limites de l'étude

La méthode proposée dans ce travail présente plusieurs limites qui doivent faire l'objet de recherches ultérieures. En effet, les situations d'agrégation, les modifications temporelles, l'utilisation sur données agrégées, la nécessité d'avoir des matrices de transition ainsi que les

limites de l'utilisation du PMSI dans le domaine de l'analyse spatiale sont autant d'obstacles qu'il est indispensable de surmonter. Ces 5 limites ainsi que leurs remédiations possibles vont être détaillées dans les sections suivantes.

A. Situations d'agrégation

La situation identifiée comme numéro 5 dans le Table 4 peut entrainer des agrégations fortes. Il s'agissait en effet de la situation d'un *Spatial_Id_INSEE* partagé entre 2 *Spatial_Id_PMSI* qui étaient fusionnés pour créer un seul et unique *Spatial_Id_analyse.* Cette situation était rare dans l'application présentée correspondant à moins de 1% des communes et moins de 1 % de la population. Néanmoins, il est important de connaître cette fréquence par une description détaillée des matrices de passage. Par ailleurs, il est nécessaire de détecter des situations d'agrégation trop fortes ou trop fréquentes. En effet la méthode a été conçue dans l'objectif de limiter l'agrégation au maximum. Si par son utilisation l'agrégation devient trop importante, l'intérêt de son utilisation diminue. Ainsi l'utilisation de l'indice de dégradation de résolution spatiale permet de mesurer l'effet d'agrégation sur l'ensemble de la zone étudiée. Un développement intéressant serait de décliner cet indice pour chaque unité spatiale initiale. La surface de chaque unité spatiale initiale pourrait être comparée à l'unité spatiale d'analyse à laquelle elle est liée. L'analyse de la distribution de ces valeurs ferait apparaître d'éventuelles valeurs aberrantes qui correspondraient à des agrégations trop fortes ou trop fréquentes. Cet indice ciblé resterait d'interprétation relative à d'autres unités administratives et non pas absolue. Si l'indice reste à des valeurs faibles en comparaison des unités administratives, alors l'agrégation est limitée. Dans les situations d'indice élevé, la table de correspondance devrait être si possible modifiée. Dans le cas où cela est impossible, le chercheur doit se poser la question de l'adaptation de l'unité spatiale trop agrégée au phénomène étudié et il sera

nécessaire d'interpréter avec précaution les résultats de l'analyse spatiale relatifs à cette unité spatiale.

B. Variation dans le temps des unités spatiales

Un obstacle régulièrement documenté dans l'analyse spatiale est la modification des limites géographiques des unités spatiales dans le temps. En France, par exemple, les cantons ont été réformés en 2014 pour voir leur nombre divisé par 2 (121). Ainsi, des données disponibles à une même unité spatiale théorique, le canton français, sur plusieurs années ne correspondront pas à d'identiques unités spatiales réelles. Les modifications des limites communales en sont un autre exemple. La table de correspondance créée à l'issue de l'application de la méthode générique ne prend pas en compte ce paramètre et suppose les unités spatiales stables dans le temps. L'utilisation de données pluri-annuelles est cependant riche d'opportunités : mise en évidence de décalage dans le temps entre exposition à un facteur écologique et la survenue d'un événement de santé, augmentation du nombre d'évènements analysables permettant d'augmenter la puissance, suivi des dynamiques dans le temps d'un évènement de santé sur un territoire. L'intégration de la problématique temporelle dans l'analyse spatiale en santé fait l'objet d'une importante littérature qui sera abordée plus loin dans ce travail tant les perspectives sont intéressantes.

C. Données agrégées

Les données de santé sont le plus souvent des données agrégées du fait de la nécessaire conservation du secret médical. Les données du NHS, celle de la base SPARCS de l'état de New York ou du PMSI en France conservent comme lieu de résidence du patient une information issue du code postal. C'est pourquoi la méthode développée s'est intéressée uniquement à ce type de données. Cependant les données écologiques et en particulier environnementales peuvent être fréquemment disponibles à leurs coordonnées géographiques précises, c'est à dire sous forme ponctuelles ou géostatistiques. Il est ainsi possible d'avoir

accès aux données de pollution à chaque station de mesure. L'utilisation de ces données nécessite une étape de traitement préalable à l'analyse spatiale permettant l'agrégation des données géostatistiques en données latticielles. Ce traitement fait appel à des méthodes qui sont du ressort du spécialiste. Dans l'exemple de la station de mesure de la pollution, les données sont disponibles aux coordonnées X/Y reparties sur le territoire. L'utilisation de modèles d'interpolation spatiale tels que le *kriegage* permet d'obtenir pour chaque point du territoire une valeur de pollution (23). En effet, ces modèles permettent à partir de l'ensemble des valeurs connues de pollution de modéliser et d'estimer les valeurs de pollution pour chaque point du territoire (considéré comme un espace continu). Une dernière étape d'agrégation des valeurs pour chaque unité spatiale d'analyse doit ensuite être réalisée. Là encore, la méthode d'agrégation retenue dépend d'un avis expert. Certains polluant pouvant être ainsi ramenés à leur médiane ou leur moyenne quand d'autres devront être ramenés au cumul de dose sur la zone. Après cette dernière étape d'agrégation, les données médicales ainsi que les données écologiques seront disponibles pour chaque unité spatiale d'analyse.

D. L'existence de matrices de transition

Une des limites de la méthode proposée est simple et peut sembler évidente mais reste un obstacle majeur. En effet, la construction de la table de correspondance suppose de pouvoir identifier une ou plusieurs matrices de transition mettant en lien le *Spatial_Id_eco* et le *Spatial_Id_medical.* Si aucune matrice de transition n'existe, la construction de la table de correspondance est impossible et il faudra utiliser d'autres méthodes pour surmonter le *change of support problem*. Cependant cette situation apparaît comme rare car les identifiants spatiaux des bases de données doivent être à la fois faciles à récupérer et sont rarement créés *de novo*. Il est peu probable qu'aucune unité administrative existante ne puisse permettre de débuter la chaîne des matrices de transition. Néanmoins, dans cette situation, une solution consiste à utiliser les systèmes d'information géographique afin de réaliser des intersections

spatiales pour faire correspondre l'unité sans matrice de transition avec l'unité spatiale finale. C'est à dire qu'il est nécessaire de représenter sous forme cartographique l'ensemble des unités spatiales des 2 bases de données puis de rechercher les limites de nouvelles unités spatiales ne fragmentant aucune des unités spatiales initiales. Cette démarche doit être réalisée avec précaution.

E. Les limites du PMSI

Les limites de l'utilisation de la base de données du PMSI dans l'analyse spatiale en santé sont de 2 ordres. Les premières difficultés sont celles développées dans la littérature pour l'épidémiologie qui ont d'ores et déjà été largement décrites que cela soit en général, pour l'épidémiologie des cancers ou pour la surveillance des infections nosocomiales (57–60). Il est possible de citer par exemple tout d'abord le changement de codage du diagnostic principal qui a eu lieu en 2009 puis les modifications annuelles de règles de codage pouvant modifier totalement l'approche de certaines pathologies comme cela a eu lieu pour le codage des sepsis. Enfin la qualité de codage dans les établissements est variable selon que le codage a un impact sur la valorisation des séjours ou non. Les secondes difficultés sont spécifiques à l'analyse spatiale et sont encore mal évaluées. Tout d'abord, si les modifications de codes géographiques PMSI sont rares et bien identifiées par l'Atih, les difficultés proviennent des codes erronés et inexistants. En effet, les données de lieu de résidence sont enregistrées dans les logiciels de Gestion Administrative du Malade (GAM) sous la forme de codes postaux. L'algorithme d'anonymisation et de groupage GenRSA transforme ces codes postaux en codes géographiques PMSI grâce à la table mise à disposition par l'ATIH. Si le code postal est inconnu, dans le cas d'un code cedex par exemple, le code géographique ne sera pas renseigné. Ensuite, si le code postal est existant mais erroné par erreur du patient, erreur de saisie de l'agent administratif ou du fait d'un problème technique alors le code géographique PMSI correspondant sera erroné. La seule remédiation possible est aujourd'hui de supprimer

les patients sans codes géographiques valides, enregistrés sous la forme d'un code « 99999 ». Néanmoins, il serait utile d'intégrer dans les logiciels de saisies des informations administratives des contrôles qualités permettant de détecter ces erreurs quand elles peuvent encore être corrigées. Ainsi un logiciel qui testerait la correspondance entre rue, ville et code postal et exécuterait systématiquement la correspondance avec le code géographique PMSI pourrait permettre de pallier sensiblement à ces difficultés.

V. Perspectives

A. La prise en compte de la notion de temps

L'intégration d'une composante temporelle dans les analyses spatiales représente une piste de développement majeure de la méthode. Le premier intérêt est de pouvoir utiliser des bases de données correspondant à plusieurs années en conservant la temporalité des évènements. En effet, pouvoir utiliser des données sur plusieurs années permet d'augmenter le nombre d'individus statistiques pris en compte dans l'analyse et d'augmenter la puissance. Le second intérêt est de pouvoir rechercher des associations entre évènement de santé et exposition à un facteur de risque décalés dans le temps ou pour lesquels les durées d'exposition jouent un rôle. Par exemple, dans la recherche d'un lien entre exposition à un facteur de risque et survenue d'un cancer, il peut souvent s'écouler une durée de plusieurs années appelée latence. Une analyse spatiale qui n'intègrerait pas des données pluri-annuelles ne pourrait pas mettre correctement en évidence cette association.

Actuellement il existe des méthodes statistiques permettant de prendre en compte l'importance du temps. Il existe par exemple des méthodes pour évaluer la durée pendant laquelle existe un cluster de cas grâce aux statistiques de scan spatio-temporelles de Kuldorff

(20). Cette méthode a par exemple été appliquée sur des données du registre Epimad par Genin et al. (122). Les modèles de type bayésien appartiennent aussi aux outils permettant d'étudier la durée de la latence entre la survenue de l'évènement de santé et l'exposition à un facteur de risque écologique comme l'a montré Knorr-Held dans une application sur les cancer du poumon dans l'Ohio, USA (123).

Cependant ces approches prometteuses se heurtent aux problèmes du *Modifiable temporal unit problem* (MTUP). Dans une étude sur les effets du MTUP sur la détection de cluster, Cheng et al. reprennent les 3 aspects du MTUP. Il y a tout d'abord l'effet d'agrégation qui consiste à choisir quelle doit être l'agrégation temporelle qui peut être le jour, le mois ou l'année. Ce choix doit se faire selon l'évènement de santé étudié. Une latence longue de l'évènement imposera le choix d'une agrégation plus forte, annuelle par exemple dans le cas du cancer. Ensuite le deuxième aspect du MTUP est celui de la segmentation des durées. Une agrégation sur 7 jours pourra par exemple être segmenté au lundi, mercredi ou dimanche. De la même façon que pour l'agrégation, le choix de la segmentation doit prendre en compte l'événement étudié. Il peut être utile par exemple de travailler sur des données annuelles mais sans coupure de l'hiver pour suivre la dynamique d'épidémies telles que la grippe saisonnière. Enfin le 3ème aspect du MTUP est celui de l'effet de frontière qui se rapporte à la définition d'un début et d'une fin dans les données analysées. La réponse à ce problème doit prendre tout d'abord en compte les données disponibles ainsi que la latence supposée entre exposition et événement. L'étude des facteurs de risques d'une pathologie telle que le diabète de type 2 sur des durées courtes n'est pas pertinente compte tenu de la dynamique conduisant à l'apparition de la maladie. Le MTUP constitue un obstacle récent des analyses spatio-temporelles qu'il est important de connaître afin de pouvoir définir des protocoles d'étude optimaux.

Une difficulté supplémentaire de l'analyse spatio-temporelle est la variation des limites géographiques dans le temps. Cela signifie qu'un découpage géographique valable à une année *n* ne l'est plus à une année antérieure ou postérieure. Une des solutions possibles à ce problème est tout d'abord de choisir l'année de référence pour la réalisation des analyses. Il sera ensuite nécessaire de transformer les données d'unités spatiales des autres années en données à l'unité spatiale de référence. Cela peut être réalisé en construisant selon la même méthode que proposée dans cette étude des tables de correspondance entre différentes années pour une unité spatiale donnée. Ainsi les unités spatiales sont lissées dans le temps bien que toutes les données médicales et écologiques restent caractérisées par un attribut temporel (date de début et de fin, date de survenue…).

L'approche temporelle est particulièrement intéressante sur le plan épidémiologique et doit faire l'objet de recherches pour que les difficultés inhérentes à l'intégration de la donnée temps puissent être surmontées.

B. L'exploitation des données françaises

Au travers du PMSI, la France dispose d'une base de données très riche des évènements de santé. En effet, cette base de données intègre l'ensemble des séjours d'hospitalisation en France. Les informations médicales (diagnostics, actes médicaux), administratives (âge, sexe…) et spatiales (lieu de résidence du patient, identification de l'établissement de santé) sont présentes dans la base de données. De plus, chaque patient est identifié par un numéro anonyme qui permet de lier entre eux les différents séjours hospitaliers. La mise à disposition de la base nationale du PMSI aux équipes de recherche est aujourd'hui aisée. Cependant, si l'exploitation épidémiologique de cette base de données est abondante, l'utilisation en analyse spatiale reste plus rare ou à de faibles résolutions spatiales telles que le département. Une des

analyses spatiales les plus aboutie à partir du PMSI est le travail de Roussot et al. sur les accidents vasculaires cérébraux qui utilise le code géographique PMSI comme unité spatiale d'analyse (62). En effet, une étude de l'Irdes sur le recours à différentes chirurgies agrégeait les données à l'échelle du département et une étude de la Drees transformait les données du PMSI en données disponibles à la commune par une imputation aléatoire (54,55).

Notre méthode permet de fusionner de façon simple et claire une ou plusieurs bases de données écologiques à l'échelle de la commune avec les données du PMSI. Ainsi il devient possible d'étudier un très grand nombre de problèmes de santé. Il peut s'agir de rechercher des facteurs de risque mais aussi d'analyser l'accès aux soins. Par ailleurs il semble intéressant de croiser les données du PMSI avec des données de registres sur l'ensemble du territoire français. Cela ouvrirait l'opportunité de valider aisément la fiabilité des données du PMSI pour un événement de santé.

Le PMSI offre de nombreuses opportunités à être utilisé dans le cadre de l'analyse spatiale sur l'ensemble du territoire national. Cependant, l'analyse spatiale en santé requière une base de données médicales mais aussi une ou plusieurs bases de données écologiques. Ces dernières sont aisément disponibles dans le domaine socio-économique grâce à l'INSEE qui met à disposition des éléments aussi variés que les données démographiques, les données relatives au niveau socio-économique ou les données sur le tissu économique français. Néanmoins, l'obtention des données environnementales sur l'ensemble du territoire français telles que le niveau de pollution atmosphérique, le niveau de pollution des nappes phréatique ou les émissions de métaux lourds reste une difficulté importante. La création de telles bases de données environnementales constitue un défi de taille. Il est cependant indispensable si l'on souhaite par exemple répliquer sur l'ensemble du territoire français des études telles que celle de Cocker et al. sur le lien entre un faible poids de naissance et exposition aux particules de moins de 2,5 micromètres (25).

Conclusion générale

La méthode développée dans cette étude a permis de créer une table de correspondance sans utiliser de méthode de désagrégation spatiale et en minimisant l'agrégation. La méthode est reproductible selon un schéma d'étude standardisé qui facilite son application à de multiples bases de données qui sont aujourd'hui de plus en plus disponibles. En particulier en France, les premières applications montrent qu'il est possible d'exploiter le PMSI ainsi que des bases des données socio-économiques de l'INSEE et le French EDI. Cela ouvre ainsi un champ d'étude particulièrement large sur l'étude de la répartition spatiale de nombreux évènements de santé et la détermination de leurs facteurs de risques qu'il convient de compléter dans les développements futurs par l'intégration de la composante temporelle.

Références

1. Bouyer J, Cordier S, Levallois P. Epidémiologie. Edisem. 2003. 89-118 p. (Environnement et santé publique - Fondements et pratiques).

2. Morgenstern H. Ecologic Studies in Epidemiology: Concepts, Principles, and Methods. Annu Rev Public Health. 1995;16(1):61–81.

3. Walter SD. The ecologic method in the study of environmental health. II. Methodologic issues and feasibility. Environ Health Perspect. 1991 Aug;94:67–73.

4. Walter SD. The ecologic method in the study of environmental health. I. Overview of the method. Environ Health Perspect. 1991 Aug;94:61–5.

5. Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, et al. Construction of an adaptable European transnational ecological deprivation index: the French version. J Epidemiol Community Health. 2012 Nov;66(11):982–9.

6. Morgenstern H. Uses of ecologic analysis in epidemiologic research. Am J Public Health. 1982 Dec;72(12):1336–44.

7. Selvin HC. Durkheim's Suicide and Problems of Empirical Research. Am J Sociol. 1958;63(6):607–19.

8. Robinson WS. Ecological Correlations and the Behavior of Individuals. Am Sociol Rev. 1950;15(3):351–7.

9. Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic Issues and Approaches to Spatial Epidemiology. Environ Health Perspect. 2008 Aug;116(8):1105–10.

10. Melnick AL, Fleming DW. Modern Geographic Information Systems- Promise and Pitfalls. J Public Health Manag Pract [Internet]. 1999;5(2). Available from: http://journals.lww.com/jphmp/Fulltext/1999/03000/Modern_Geographic_Information_Systems__Promise_and.3.aspx

11. Lee LP. The Transverse Mercator Projection of the Spheroid. Emp Surv Rev. 1945

Oct 1;8(58):142–52.

12. Snyder JP. Map projections used by the U.S. Geological Survey [Internet]. Washington, D.C.: U.S. Government Printing Office; 1982 [cited 2017 Jun 22]. (Bulletin). Report No.: 1532. Available from: http://pubs.er.usgs.gov/publication/b1532

13. Snow J. On the Mode of Communication of Cholera. John Churchill; 1855. 216 p.

14. Howard-Jones N. Robert Koch and the cholera vibrio: a centenary. Br Med J Clin Res Ed. 1984 Feb 4;288(6414):379–81.

15. Nardi MG. [Discovery of Vibrio cholerae by Filippo Pacini, of Pistoia, established in the initial phases of microbiological thought and judged after a century]. Minerva Med. 1954 Dec 22;45(102):Varia, 1024–9.

16. Environmental Systems Research Institute. ArcGIS. Esri; 2011.

17. QGIS Development Team. QGIS Geographic Information System. Open Source Geospatial Foundation; 2009.

18. R Core Team. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.R-project.org/

19. Clarke KC, McLafferty SL, Tempalski BJ. On epidemiology and geographic information systems: a review and discussion of future directions. Emerg Infect Dis. 1996 Jun;2(2):85–92.

20. Kulldorff M. SaTScan: Software for the spatial and space-time scan statistics.

21. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Stat Comput. 2000 Oct 1;10(4):325–37.

22. Graham AJ, Atkinson PM, Danson FM. Spatial analysis for epidemiology. Acta Trop. 2004 Aug 1;91(3):219–25.

23. OLIVER MA, WEBSTER R. Kriging: a method of interpolation for geographical

information systems. Int J Geogr Inf Syst. 1990 Jul 1;4(3):313–32.

24. Joly D, Thierry B, Cardot H, Cavailhès J, Hilal M, Wavresky P. Interpolation par régressions locales : application aux précipitations en France, Résumé, Interpolation by local regressions applied to precipitation in FranceAbstract. L'Espace Géographique. 2009 Jun 5;38(2):157–70.

25. Coker E, Ghosh J, Jerrett M, Gomez-Rubio V, Beckerman B, Cockburn M, et al. Modeling spatial effects of PM2.5 on term low birth weight in Los Angeles County. Environ Res. 2015 Oct;142:354–64.

26. Atmo. Programme régional de surveillance de la qualité de l'air 2017 - 2021 des Hauts de France. 2017.

27. Genin M. Statistiques de scan : théorie et application à l'épidémiologie. [Internet]. [Lille]: Université du droit et de la santé - Lille 2; 2013. Available from: https://tel.archives-ouvertes.fr/tel-01004929/document

28. Scott L, Mobley LR, Il'yasova D. Geospatial Analysis of Inflammatory Breast Cancer and Associated Community Characteristics in the United States. Int J Environ Res Public Health. 2017 Apr 11;14(4).

29. Takahashi K, Tachimori H, Kan C, Nishi D, Okumura Y, Kato N, et al. Spatial analysis for regional behavior of patients with mental disorders in Japan. Psychiatry Clin Neurosci. 2017 Apr;71(4):254–61.

30. Sánchez-Díaz G, Arias-Merino G, Villaverde-Hueso A, Morales-Piga A, Abaitua-Borda I, Hens M, et al. Monitoring Huntington's Disease Mortality across a 30-Year Period: Geographic and Temporal Patterns. Neuroepidemiology. 2016 Nov 25;47(3-4):155–63.

31. Nicholl J, Jacques RM, Campbell MJ. Direct risk standardisation: a new method for comparing casemix adjusted event rates using complex models. BMC Med Res Methodol. 2013 Oct 29;13:133.

32. Inskip H, Beral V, Fraser P, Haskey J. Methods for age-adjustment of rates. Stat Med. 1983 Dec;2(4):455–66.

33. Declercq C, Gower-Rousseau C, Vernier-Massouille G, Salleron J, Baldé M, Poirier G, et al. Mapping of inflammatory bowel disease in northern France: spatial variations and relation to affluence. Inflamm Bowel Dis. 2010 May;16(5):807–12.

34. Kulldorff M, Tango T, Park PJ. Power comparisons for disease clustering tests. Comput Stat Data Anal. 2003 Apr 28;42(4):665–84.

35. Goujon-Bellec S, Demoury C, Guyot-Goubin A, Hémon D, Clavel J. Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. Int J Health Geogr. 2011 Oct 4;10:53.

36. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med. 1995;14(8):799–810.

37. Kulldorff M. A spatial scan statistic. Commun Stat - Theory Methods. 1997 Jan 1;26(6):1481–96.

38. Zhang Z, Assunção R, Kulldorff M. Spatial Scan Statistics Adjusted for Multiple Clusters. J Probab Stat. 2010;2010:e642379.

39. Kulldorff M, Athas WF, Feurer EJ, Miller BA, Key CR. Evaluating cluster alarms: a space-time scan statistic and brain cancer in Los Alamos, New Mexico. Am J Public Health. 1998 Sep;88(9):1377–80.

40. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Stat Math. 1991 Mar 1;43(1):1–20.

41. Auchincloss AH, Gebreab SY, Mair C, Roux AVD. A Review of Spatial Methods in Epidemiology, 2000–2010. Annu Rev Public Health. 2012 Apr;33:107–22.

42. Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ Geogr. 1970;46:234–40.

43. Lawson AB. Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Second Edition. CRC Press; 2013. 398 p.

44. Rooney J, Vajda A, Heverin M, Crampsie A, Tobin K, McLaughlin R, et al. No association between soil constituents and amyotrophic lateral sclerosis relative risk in Ireland. Environ Res. 2016 May;147:102–7.

45. Goria S, Stempfelet M, de Crouy-Chanel P. Introduction aux statistiques spatiales et aux systèmes d'information géographique en santé environnement, Application aux études écologiques. [Internet]. Institut de Veille Sanitaire. Saint-Maurice; 2011. 65 p. Available from: http://www.dphu.org/uploads/attachements/books/books_355_0.pdf

46. Gotway CA, Young LJ. Combining Incompatible Spatial Data. J Am Stat Assoc. 2002 Jun 1;97(458):632–48.

47. Nerich V, Monnet E, Etienne A, Louafi S, Ramée C, Rican S, et al. Geographical variations of inflammatory bowel disease in France: a study based on national health insurance data. Inflamm Bowel Dis. 2006 Mar;12(3):218–26.

48. Bureau of Health Informatics Office of Quality and Patient Safety. SPARCS Operations Guide [Internet]. NYS Department of Health; 2016 [cited 2016 Dec 16]. Available from: https://www.health.ny.gov/statistics/sparcs/training/docs/sparcs_operations_guide.pdf

49. Agence technique de l'information hospitalière. Guide méthodologique de production des informations relatives à l'activité médicale et àa sa facturation en médecine, chirurgie, obstétrique et odontologie. Ministère des affaires sociales et de la santé. 2016. (Bulletin officiel).

50. Health and social care information centre. HES Data Dictionary: Admitted Patient Care [Internet]. NHS Digital; 2017 [cited 2017 Mar 10]. Available from: http://content.digital.nhs.uk/media/23711/Admitted-Patient-Care/pdf/Admitted_Patient_Care_.pdf

51. A. Kadadi, R. Agrawal, C. Nyamful, R. Atiq. Challenges of data integration and interoperability in big data. In: 2014 IEEE International Conference on Big Data (Big Data). 2014. p. 38–40.

52. Toga AW, Foster I, Kesselman C, Madduri R, Chard K, Deutsch EW, et al. Big biomedical data as the key resource for discovery science. J Am Med Inform Assoc JAMIA. 2015 Nov;22(6):1126–31.

53. Jones SG, Kulldorff M. Influence of spatial resolution on space-time disease cluster detection. PloS One. 2012;7(10):e48036.

54. Le Bail M, Or Z. Atlas des variations de pratiques médicales. Recours à dix interventions chirurgicales. [Internet]. Irdes; 2016. Available from: www.irdes.fr/recherche/ouvrages/002-atlas-des-variations-de-pratiques-medicales-recours-a-dix-interventions-chirurgicales.pdf

55. Evain F. À quelle distance de chez soi se fait-on hospitaliser ? Etudes Résultats Drees [Internet]. 2011 fevrier [cited 2016 Dec 15];(754). Available from: http://drees.social-sante.gouv.fr/IMG/pdf/er754-2.pdf

56. Bocquier A, Thomas N, Zitouni J, Lewandowski E, Cortaredona S, Jardin M, et al. Évaluation de la qualité du chaînage des séjours hospitaliers pour l'étude des variations spatiales de santé à partir des données du PMSI. Étude de faisabilité dans trois régions françaises. Rev DÉpidémiologie Santé Publique. 2011 Aug;59(4):243–9.

57. Carré N, Uhry Z, Velten M, Trétarre B, Schvartz C, Molinié F, et al. [Predictive value and sensibility of hospital discharge system (PMSI) compared to cancer registries for thyroïd cancer (1999-2000)]. Rev Epidemiol Sante Publique. 2006 Sep;54(4):367–76.

58. COURIS C-M, ECOCHARD R. Utilisation du PMSI pour l'épidémiologie. Faisabilité, conditions d'utilisation et limites. Tarif À Act. 2005 Oct;(449):651–4.

59. Gerbier S, Bouzbid S, Pradat E, Baulieux J, Lepape A, Berland M, et al. [Use of the

French medico-administrative database (PMSI) to detect nosocomial infections in the University hospital of Lyon]. Rev Epidemiol Sante Publique. 2011 Feb;59(1):3–14.

60. Olive F, Gomez F, Schott A-M, Remontet L, Bossard N, Mitton N, et al. Analyse critique des données du PMSI pour l'épidémiologie des cancers : une approche longitudinale devient possible. /data/revues/03987620/v59i1/S0398762010005018/ [Internet]. 2011 Oct 2 [cited 2016 Dec 15]; Available from: http://www.em-consulte.com/en/article/281036

61. Coldefy M, Com-Ruelle L, Lucas-Gabrielli V, Marcoux L. Les distances d'accès aux soins en France métropolitaine au 1er janvier 2007: IRDES2011.

62. Roussot A, Cottenet J, Gadreau M, Giroud M, Béjot Y, Quantin C. The use of national administrative data to describe the spatial distribution of in-hospital mortality following stroke in France, 2008–2011. Int J Health Geogr. 2016;15(1):2.

63. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique. 2017 Jul 26;

64. Moulis G, Lapeyre-Mestre M, Palmaro A, Pugnet G, Montastruc J-L, Sailler L. French health insurance databases: What interest for medical research? Rev Med Interne. 2015 Jun;36(6):411–7.

65. Townsend P, Phillimore P, Beattie A. Health and Deprivation: Inequality and the North. Croom Helm; 1988. 250 p.

66. Briggs D. The role of GIS: Coping with space (and time) in air pollution exposure assessment. J Toxicol Environ Health A [Internet]. 2005;68. Available from: http://dx.doi.org/10.1080/15287390590936094

67. Jerret M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. J Expo Environ Epidemiol

[Internet]. 2005;15. Available from: http://dx.doi.org/10.1038/sj.jea.7500388

68. Pénard-Morand C, Schillinger C, Armengaud A, Debotte G, Chrétien E, Pellier S, et al. Assessment of schoolchildren's exposure to traffic related air pollution in the French Sic Cities Study using a dispersion model. Atmos Environ [Internet]. 2006;40. Available from: http://dx.doi.org/10.1016/j.atmosenv.2005.11.057

69. Kumarathilaka P, Oze C, Indraratne SP, Vithanage M. Perchlorate as an emerging contaminant in soil, water and food. Chemosphere. 2016 May;150:667–77.

70. Leung AM, Pearce EN, Braverman LE. Perchlorate, iodine and the thyroid. Best Pract Res Clin Endocrinol Metab. 2010 Feb;24(1):133–41.

71. Loi n° 96-1236 du 30 décembre 1996 sur l'air et l'utilisation rationnelle de l'énergie.

72. Tatem AJ, Jia P, Ordanovich D, Falkner M, Huang Z, Howes R, et al. The geography of imported malaria to non-endemic countries: a meta-analysis of nationally reported statistics. Lancet Infect Dis. 2017 Jan 1;17(1):98–107.

73. Burke M, Heft-Neal S, Bendavid E. Sources of variation in under-5 mortality across sub-Saharan Africa: a spatial analysis. Lancet Glob Health. 2016 Dec;4(12):e936–45.

74. Duncan EW, White NM, Mengersen K. Bayesian spatiotemporal modelling for identifying unusual and unstable trends in mammography utilisation. BMJ Open. 2016 May 26;6(5):e010253.

75. Vine MF, Degnan D, Hanchette C. Geographic information systems: their use in environmental epidemiologic research. Environ Health Perspect. 1997 Jun;105(6):598–605.

76. Gower-Rousseau C. Epidémiologie des maladies inflammatoires chroniques de l'Intestin en France : apport du registre EPIMAD [Internet] [phdthesis]. Université du Droit et de la Santé - Lille II; 2012 [cited 2016 Dec 20]. Available from: https://tel.archives-ouvertes.fr/tel-00820631/document

77. Bambhroliya AB, Burau KD, Sexton K. Spatial analysis of county-level breast cancer

mortality in Texas. J Environ Public Health. 2012;2012:959343.

78. Meliker JR, Jacquez GM, Goovaerts P, Copeland G, Yassine M. Spatial cluster analysis of early stage breast cancer: a method for public health practice using cancer registry data. Cancer Causes Control CCC. 2009 Sep;20(7):1061–9.

79. Goungounga JA, Gaudart J, Colonna M, Giorgi R. Impact of socioeconomic inequalities on geographic disparities in cancer incidence: comparison of methods for spatial disease mapping. BMC Med Res Methodol. 2016 Oct 12;16(1):136.

80. Yildirim O, Gottwald M, Schüler P, Michel MC. Opportunities and Challenges for Drug Development: Public-Private Partnerships, Adaptive Designs and Big Data. Front Pharmacol. 2016;7:461.

81. Barrett MA, Humblet O, Hiatt RA, Adler NE. Big Data and Disease Prevention: From Quantified Self to Quantified Communities. Big Data. 2013 Sep;1(3):168–75.

82. Sparks R, Lau WW, Tsang JS. Expanding the Immunology Toolbox: Embracing Public-Data Reuse and Crowdsourcing. Immunity. 2016 Dec 20;45(6):1191–204.

83. Marco-Ruiz L, Moner D, Maldonado JA, Kolstrup N, Bellika JG. Archetype-based data warehouse environment to enable the reuse of electronic health record data. Int J Med Inf. 2015 Sep;84(9):702–14.

84. Fletcher-Lartey SM, Caprarelli G. Application of GIS technology in public health: successes and challenges. Parasitology. 2016 Apr;143(4):401–15.

85. Ribeiro AI, Olhero A, Teixeira H, Magalhães A, Pina MF. Tools for address georeferencing - limitations and opportunities every public health professional should be aware of. PloS One. 2014;9(12):e114130.

86. Beuscart J-B, Genin M, Dupont C, Verloop D, Duhamel A, Defebvre M-M, et al. Potentially inappropriate medication prescribing is associated with socioeconomic factors: a spatial analysis in the French Nord-Pas-de-Calais Region. Age Ageing. 2017 Jan 6;

87. Devogele T, Parent C, Spaccapietra S. On spatial database integration. Int J Geogr Inf Sci. 1998 Jun 1;12(4):335–52.

88. Canto MT, Anderson WF, Brawley O. Geographic Variation in Breast Cancer Mortality for White and Black Women: 1986–1995. CA Cancer J Clin. 2001 Nov 1;51(6):367–70.

89. Elliott P, Wartenberg D. Spatial Epidemiology: Current Approaches and Future Challenges. Environ Health Perspect. 2004 Jun;112(9):998–1006.

90. Li T, Pullar DV, Corcoran J, Stimson RJ. A comparison of spatial disaggregation techniques as applied to population estimation for south east Queensland (SEQ), Australia. Appl GIS. 2007 Sep 1;3(9):1–16.

91. Spatial Disaggregation & Small-Area Estimation Methods for Agri. Surveys: Solutions & Perspectives [Internet]. Global strategy : improving agricultural and rural statistics; 2015. Available from: http://gsars.org/wp-content/uploads/2015/09/TR-Spatial-Disaggregation-and-Small-Area-Estimation-210915.pdf

92. Grubesic TH, Matisziw TC. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. Int J Health Geogr. 2006 Dec 13;5:58.

93. Siordia C, Wunneburger DF. Contiguity principle for geographic units: Evidence on the quantity, degree, and location of Public Use Microdata Area (PUMA) fragmentation. Hum Geogr - J Stud Res Hum Geogr. 2013 Nov 25;7(2):5–13.

94. Siordia C, Fox A. Public Use Microdata Area Fragmentation: Research and Policy Implications of Polygon Discontiguity. Spat Demogr. 2013 Apr 1;1(1):41–55.

95. Institut national de la statistique et des études économiques [Internet]. Available from: https://www.insee.fr

96. Open data soft. Correspondance Code INSEE - Code Postal 2013 [Internet]. [cited

2016 Dec 15]. Available from: https://public.opendatasoft.com/explore/dataset/correspondance-code-insee-code-postal/

97. Atih. Mise à jour 2014 de la liste de correspondance codes postaux codes géographiques [Internet]. 2015 [cited 2016 Dec 16]. Available from: http://www.atih.sante.fr/mise-jour-2014-de-la-liste-de-correspondance-codes-postaux-codes-geographiques

98. Atih. Mise à jour 2015 de la liste de correspondance codes postaux codes géographiques PMSI [Internet]. 2015 [cited 2016 Dec 16]. Available from: http://www.atih.sante.fr/mise-jour-2015-de-la-liste-de-correspondance-codes-postaux-codes-geographiques

99. Rossheim ME, Thombs DL, Wagenaar AC, Xuan Z, Aryal S. High Alcohol Concentration Products Associated With Poverty and State Alcohol Policies. Am J Public Health. 2015 Sep;105(9):1886–92.

100. Sundmacher L, Busse R. The impact of physician supply on avoidable cancer deaths in Germany. A spatial analysis. Health Policy Amst Neth. 2011 Nov;103(1):53–62.

101. Olson KL, Grannis SJ, Mandl KD. Privacy Protection Versus Cluster Detection in Spatial Epidemiology. Am J Public Health. 2006 Nov;96(11):2002–8.

102. Lee J, Alnasrallah M, Wong D, Beaird H, Logue E. Impacts of Scale on Geographic Analysis of Health Data: An Example of Obesity Prevalence. ISPRS Int J Geo-Inf. 2014 Oct 24;3(4):1198–210.

103. Jeffery C, Ozonoff A, Pagano M. The effect of spatial aggregation on performance when mapping a risk of disease. Int J Health Geogr. 2014 Mar 13;13:9.

104. Gelfand AE, Zhu L, Carlin BP. On the change of support problem for spatio-temporal data. Biostat Oxf Engl. 2001 Mar;2(1):31–45.

105. Quantin C, Cottenet J, Vuagnat A, Prunet C, Mouquet M-C, Fresson J, et al. Qualité

des données périnatales issues du PMSI : comparaison avec l'état civil et l'enquête nationale périnatale 2010. J Gynécologie Obstétrique Biol Reprod. 2014 Nov 1;43(9):680–90.

106. Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M, et al. Development of a cross-cultural deprivation index in five European countries. J Epidemiol Community Health. 2016 May;70(5):493–9.

107. Sánchez-Santos MT, Mesa-Frias M, Choi M, Nüesch E, Asunsolo-Del Barco A, Amuzu A, et al. Area-level deprivation and overall and cause-specific mortality: 12 years' observation on British women and systematic review of prospective studies. PloS One. 2013;8(9):e72656.

108. Gillespie LD, Gillespie WJ, Robertson MC, Lamb SE, Cumming RG, Rowe BH. Interventions for preventing falls in elderly people. Cochrane Database Syst Rev. 2003;(4):CD000340.

109. Malavolta N, Rossi E, Buffa A, Falchetti A. Fragility fractures: clinical and therapeutic aspects. J Biol Regul Homeost Agents. 2015 Dec;29(4):761–9.

110. Petis S, Howard JL, Lanting BL, Vasarhelyi EM. Surgical approach in primary total hip arthroplasty: anatomy, technique and clinical outcomes. Can J Surg J Can Chir. 2015 Apr;58(2):128–39.

111. Cong Y, Zhao J, Bao N, Zeng X, Guo T, Cheng X, et al. [Prognostic significance of hidden blood loss in total hip arthroplasty (THA)]. Zhongguo Gu Shang China J Orthop Traumatol. 2011 Jun;24(6):466–8.

112. Van Kasteren MEE, Manniën J, Ott A, Kullberg B-J, de Boer AS, Gyssens IC. Antibiotic prophylaxis and the risk of surgical site infections following total hip arthroplasty: timely administration is the most important factor. Clin Infect Dis Off Publ Infect Dis Soc Am. 2007 Apr 1;44(7):921–7.

113. Lavernia CJ, Iacobelli DA, Brooks L, Villa JM. The Cost-Utility of Total Hip

Arthroplasty: Earlier Intervention, Improved Economics. J Arthroplasty. 2015 Jun;30(6):945–9.

114. Crandall CJ, Han W, Greendale GA, Seeman T, Tepper P, Thurston R, et al. Socioeconomic status in relation to incident fracture risk in the Study of Women's Health Across the Nation. Osteoporos Int J Establ Result Coop Eur Found Osteoporos Natl Osteoporos Found USA. 2014 Apr;25(4):1379–88.

115. Marks R. Hip fracture epidemiological trends, outcomes, and risk factors, 1970–2009. Int J Gen Med. 2010 Apr 8;3:1–17.

116. Audirac P-A. Les personnes âgées, de la vie de famille à l'isolement. Econ Stat. 1985;175(1):39–54.

117. PITAUD P. Solitude et isolement des personnes âgées. Eres; 2013. 228 p.

118. La Poste, IGN. Base Adresse Nationale [Internet]. 2017. Available from: https://adresse.data.gouv.fr

119. OpenStreetMap contributors. OpenStreet Map [Internet]. 2017. Available from: https://www.openstreetmap.org/

120. Ozonoff A, Jeffery C, Manjourides J, White LF, Pagano M. Effect of spatial resolution on cluster detection: a simulation study. Int J Health Geogr. 2007;6(1):52.

121. LOI n° 2013-403 du 17 mai 2013 relative à l'élection des conseillers départementaux, des conseillers municipaux et des conseillers communautaires, et modifiant le calendrier électoral. 2013-403 mai, 2013.

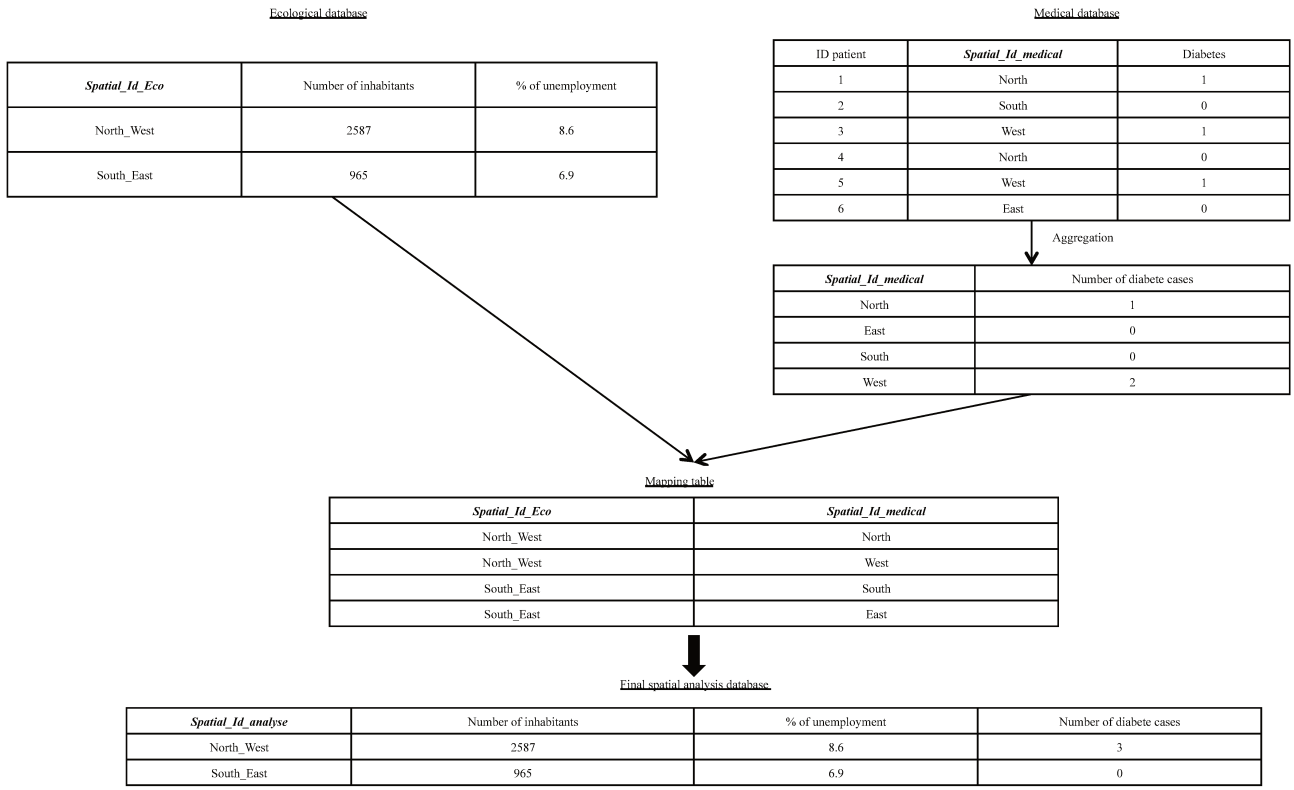
122. Genin M, Duhamel A, Preda C, Fumery M, Savoye G, Peyrin-Biroulet L, et al. Space-time clusters of Crohn's disease in northern France. J Public Health. 2013;21(6):497–504.

123. Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Stat Med. 2000 Sep 15;19(17-18):2555–67.

Annexes

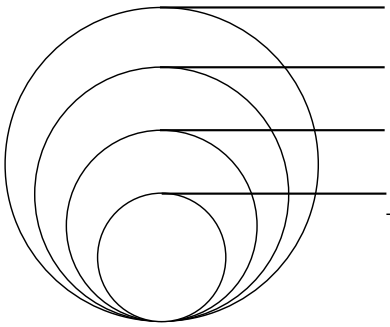
I. Additional file 1

Illustrative example of the method for building a final database for spatial analysis.



II. Additional file 2

Description of the French administrative spatial units in terms of frequencies, surface area (in km^2) and number of inhabitants. The different circles indicate the hierarchical relationships between the different administrative units.



Administrative area	Count	Median surface area (km^2)	Median number of inhabitants
Régions	18	23,669	2,136,100
Départements	96	5,987	540,900
Cantons	4,055	146	10,400
Communes	36,594	11	400

III. Additional file 3

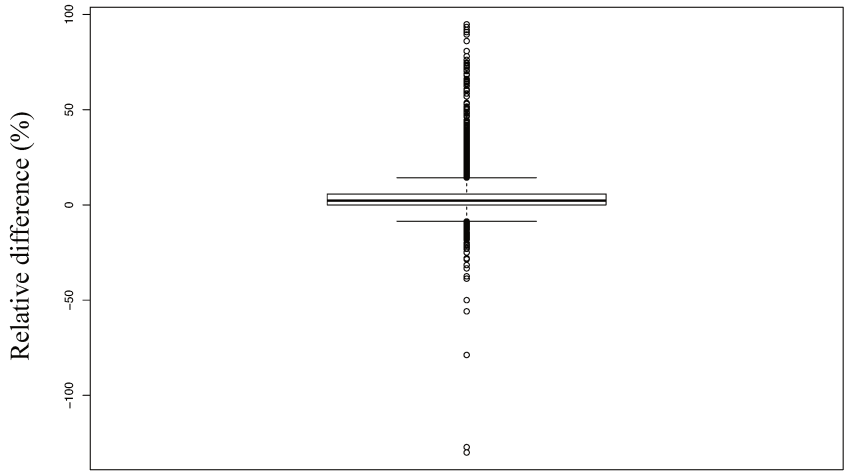
Equivalence situations for the transition matrices M_1 and M_2.

Situations	Matrix 1		Matrix 2	
	Spatial_Id_INSEE ⟷	*Zip_Code*	*Zip_Code* ⟷	*Spatial_Id_PMSI*
1 ⟷ 1	Yes		Yes	
n ⟷ 1	Yes		Yes	
1 ⟷ n	Yes		No	
n ⟷ n	No		No	

Legend: Matrix 1 is the tool used to link the Spatial_Id_INSEE and the Zip_Code. Matrix 2 is the tool used to link the Zip_Code and the Spatial_Id_PMSI. “Yes” indicates situations encountered in the application.

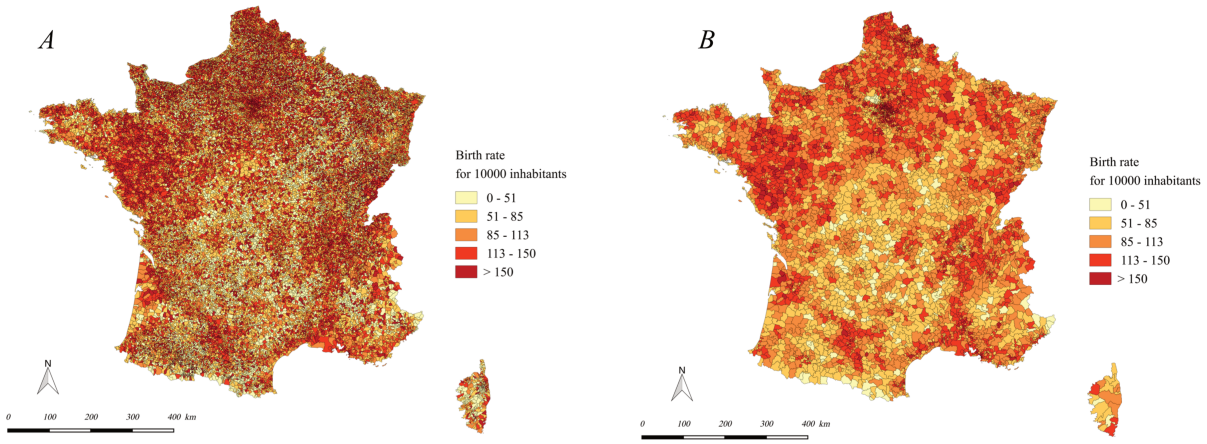
IV. Additional file 4

Relative difference in the number of births per spatial unit *SU_analysis* between the ecological data from the INSEE and the medical data from the PMSI.



V. Additional file 5

Birth rate for 10,000 inhabitants for the *SU_INSEE* (A) and *SU_analysis* (B).



Legend: The first map (A) is on the scale of the *commune* spatial unit (*SU_INSEE*), and represents the birth rate calculated from the number of births and the underlying population data in the INSEE database. The second map (B) is on the scale of the *SU_analysis* spatial unit, and represents the birth rate per 10,000 inhabitants calculated using the number of births in the PMSI database and the underlying population data in the INSEE database.

VI. Annexe 6

Codes diagnostics CIM 10 d'inclusion de Fracture de l'extrémité supérieur du fémur.

S72.0	fracture du col du fémur
S72.1	fracture du trochanter
S72.2	Fracture sous-trochantérienne
S72.9	Fracture du fémur, partie non précisée
S72.00	fracture fermée du col du fémur
S72.01	fracture ouverte du col du fémur
S72.10	Fracture fermée du trochanter
S72.11	Fracture ouverte du trochanter
S72.20	Fracture fermée sous trochantérienne
S72.21	Fracture ouverte sous trochantérienne

VII. Annexe 7

Codes diagnostics CIM 10 d'exclusion de Fracture de l'extrémité supérieur du fémur au moment du diagnostic.

M966 Fracture de hanche sur prothèse

T84 Complications de prothèses, implants et greffes orthopédiques internes

T840 Complication mécanique d'une prothèse articulaire interne

T841 Complication mécanique d'une prothèse interne de fixation d'os d'un membre

T842 Complication mécanique d'une prothèse interne de fixation d'autres os

T843 Complication mécanique d'autres prothèses, implants et greffes des os

T844 Complication mécanique d'autres prothèses, implants et greffes orthopédiques internes

T845 Infection et réaction inflammatoire dues à une prothèse articulaire interne

T846 Infection et réaction inflammatoire dues à un appareil de fixation interne [toute localisation] T847 Infection et réaction inflammatoire dues à d'autres prothèses, implants et greffes orthopédiques internes

T848 Autres complications de prothèses, implants et greffes orthopédiques internes

T849 Complication d'une prothèse, d'un implant et d'une greffe orthopédiques internes, sans précision

M00 Arthrites à bactéries pyogènes

M01 Arthrites infectieuses directes au cours de maladies infectieuses et parasitaires classées ailleurs

T029 Fractures multiples

Codes diagnostics CIM 10 d'exclusion de Fracture de l'extrémité supérieur du fémurdans les antécédents (y compris le séjour d'inclusion)

C40 Tumeur maligne des os et du cartilage articulaire des membres

C40.2 Tumeur maligne des os longs du Mb inf

C40.9 Tumeur maligne des os et du cartilage articulaire d'un membre, sans précision

C79.5 Tumeur maligne secondaire des os et de la moelle osseuse

C900 Myélome multiple

M844 Fracture pathologique

M8440 Fracture pathologique, non classée ailleurs - siège multiple

M8445 Fracture pathologique, non classée ailleurs - région pelvienne et cuisse

M8448 Fracture pathologique, non classée ailleurs - autres localisations

M8449 Fracture pathologique, non classée ailleurs - siège non précisé

AUTEUR : GHENASSIA Adrien
Date de Soutenance : 20 septembre 2017
Titre de la Thèse : Méthode générique d'amélioration de l'interopérabilité des bases de données médicales et écologiques et application sur données françaises.
Thèse - Médecine - Lille 2017
Cadre de classement : Santé Publique
DES + spécialité : Santé Publique et Médecine Sociale
Mots-clés : Analyse spatiale, réutilisation de données, change of support problem, interopérabilité

Résumé :

Introduction

La disponibilité de base de données de grandes dimensions et le développement important de la réutilisation de données et du géoréférencement ont ouvert des perspectives dans l'analyse spatiale en santé. Cependant, les études à fine échelles avec des données écologiques et médicales sont limitées par le *change of support problem* et le manque d'interopérabilité entre unités spatiales. Les méthodes de désagrégation spatiale pour résoudre ce problème introduisent des erreurs dans l'estimation spatiale. De plus, il existe un manque de cadre d'étude dans ces situations. Nous présentons ici une méthode générique d'agrégation à 2 étapes permettant de fusionner des bases de données médicales et écologiques avec un schéma standard sans utiliser de modèles de désagrégation spatiale tout en maximisant la résolution spatiale.

Méthode

Premièrement une table de correspondance est construite après identification d'une ou plusieurs matrices de passage. Cette table relie les unités spatiales des bases de données originales avec les unités spatiales de la base de données finale. Deuxièmement, la table de correspondance est validée par la comparaison de variables contenues dans les 2 bases de données initiales et la vérification de la validité spatiale avec un critère de continuité spatiale et un critère de résolution spatiale.

Résultats

Nous avons utilisé cette méthode pour fusionner une base de données médicales (le programme médicalisé des systèmes d'information contenant 5644 unités spatiales) avec une base de données écologiques (issue de l'institut national de la statistique et des études économiques contenant 36594 unités spatiales). La table de correspondance finale aboutit à 5632 unités spatiales finales. Elle a été validée par la comparaison du nombre de naissances dans la base de données médicales et dans la base de données écologiques pour chaque unité spatiale finale. La médiane [intervalle inter-quartile] de la différence relative était de 2,3% [0 ; 5,7]. Le critère de continuité spatiale était faible (2,4%) et l'indice de résolution spatiale était meilleur que pour la plupart des unités administratives françaises.

Conclusions

Notre approche innovant améliore l'interopérabilité entre bases de données médicales et écologiques et facilite l'analyse spatiale à fine échelle. Nous avons montré que les modèles de désagrégation et les larges agrégations n'étaient pas nécessairement les meilleures façons de répondre au *change of support problem*.

Composition du Jury :

Président :	Monsieur le Professeur Alain Duhamel
Assesseurs :	Madame le Professeur Florence Richard
	Madame le Docteur Marie Hélène Metzger
	Monsieur le Docteur Emmanuel Chazard
	Monsieur Michaël Genin