Méthodes d'analyse de données et modèles bayésiens appliqués au contexte des inégalités socio-territoriales de santé et des expositions environnementales

Benoît Lalloué

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université de Lorraine

en Mathématiques

par

Benoît Lalloué

Méthodes d'analyse de données et modèles bayésiens appliqués au contexte des inégalités socio-territoriales de santé et des expositions environnementales

---

Ecole doctorale : Informatique, Automatique, Électronique - Électrotechnique et Mathématiques

D. F. D. : Mathématiques

Réseau doctoral de l'École des Hautes Études en Santé Publique

Equipes d'accueil : Intitut Élie Cartan de Nancy & École des Hautes Études en Santé Publique

Soutenue publiquement le *6 décembre 2013* à Nancy, devant le jury composé de :

M. Pierre Vallois	Professeur, Université de Lorraine	Président
M. Denis Hémon	DR1, Inserm	Rapporteur
M. François Husson	Professeur, Agrocampus Ouest	Rapporteur
M. Denis Zmirou-Navier	Professeur, Université de Lorraine, EHESP	Examinateur
Mme Séverine Deguen	Enseignante-Chercheur, EHESP	Directrice de thèse
M. Jean-Marie Monnez	Professeur, Université de Lorraine	Directeur de thèse

Remerciements

Je tiens tout d'abord à remercier vivement mes deux directeurs de thèse pour avoir accepté de me suivre et de me guider tout au long de ces trois années doctorales.

Merci à Séverine Deguen pour l'opportunité qu'elle m'a offerte de rejoindre l'EHESP et son projet Equit'Area avec la thématique passionnante des inégalités sociales et environnementales de santé. Je la remercie chaleureusement pour son soutien sans faille, aussi bien dans la recherche proprement dite que dans les méandres administratifs, et pour sa très grande disponibilité.

Merci également à Jean-Marie Monnez d'avoir accepté de m'encadrer durant ces trois années dans des conditions géographiques si particulières. Je le remercie profondément pour ses nombreux conseils, pour nos discussions et nos échanges qui ont apporté beaucoup de rigueur à mon travail et m'ont systématiquement permis d'améliorer la vision statistique que j'avais de ce labeur ancré dans la santé publique et l'épidémiologie.

J'exprime ma profonde gratitude à Denis Zmirou-Navier qui m'a entraîné dans « l'aventure » de la santé publique durant mon stage de Master et ensuite en me mettant en contact avec Séverine Deguen. Merci encore pour son intérêt et sa contribution à mon travail ainsi que pour avoir réduit pas mal de difficultés tout au long de ces années.

Merci à Denis Hémon et François Husson pour leur implication dans l'évaluation de ce mémoire et pour avoir accepté d'être les rapporteurs de cette thèse. Merci aussi à Pierre Vallois pour avoir accepté d'examiner ce travail mais aussi pour ses enseignements lors de mon Master.

J'ai passé de bonnes années à l'EHESP, j'y ai travaillé dans de bonnes conditions et je voudrais maintenant dire ma reconnaissance à tous ceux qui y ont contribué.

Je remercie les différents directeurs du département d'Épidémiologie et Biostatistiques (Avner Bar-Hen, Viviane Kovess, Pascal Astagneau) de l'École des Hautes Études en Santé Publique de m'avoir accueilli dans leurs locaux et de m'avoir permis d'effectuer cette thèse dans de très bonnes conditions logistiques et matérielles.

Mes remerciements sincères à tous les membres de ce département (notamment Fei, Nolwenn, Sahar et Marie-Aude) pour leur accueil, leur bonne humeur et leur sympathie, mais aussi leur soutien moral et scientifique.

Merci aussi aux différentes personnes avec qui j'ai partagé un bureau (Elisabeth, Claire, Laure, Emma, Sabine), parfois quelques mois, parfois plus d'un an, pour nos discussions, sérieuses ou non, et les nombreux rires qu'elles ont pu occasionner. Notre liste des restaurants rennais n'a malheureusement pas été entièrement complétée, mais ce n'est que partie remise !

Je souhaite également remercier l'ensemble des membres et partenaires du projet Equit'-Area. En particulier Cindy, avec qui j'ai travaillé avec grand plaisir en menant ma thèse en parallèle à la sienne, et Wahida, pour sa disponibilité et ses explications toujours claires dans des domaines dont je ne connaissais rien. Merci également à Olivier Blanchard pour son aide et son expertise sur les différentes expositions environnementales.

Merci aussi à la direction du Réseau Doctoral de l'EHESP (Olivier Thomas, Marie-Aline Bloch, mais aussi Sarah Kitar et Maud Subtil) pour leur soutien matériel et logistique. Et également à tous les doctorants avec qui j'ai pu échanger lors d'occasions aussi bien formelles qu'informelles. En particulier, merci aux membres de l'association PH'Doc et notamment à Anne-Laure, Stéphanie, Sophie, Anne-Lise, Vincent, Youssef et Clément.

Je remercie également les membres de l'Institut Élie Cartan qui m'ont toujours fait bon accueil lors de mes passages en son sein.

Je pense aussi à tous ceux que je n'ai pas mentionnés ici mais qui, de près ou de loin, m'ont aidé à effectuer cette thèse.

Pour finir, je remercie ma famille et mes amis pour tout leur soutien, leur aide et leurs encouragements, malgré les nombreux kilomètres qui nous ont séparés durant ces trois ans.

Résumés et Laboratoires

Méthodes d'analyse de données et modèles bayésiens appliqués au contexte des inégalités socio-territoriales de santé et des expositions environnementales

Cette thèse a pour but d'améliorer et d'appliquer les connaissances concernant les techniques d'analyse de données et certains modèles bayésiens dans le domaine de l'étude des inégalités sociales et environnementales de santé. À l'échelle géographique de l'IRIS sur les agglomérations de Paris, Marseille, Lyon et Lille, l'événement sanitaire étudié est la mortalité infantile dont on cherchera à expliquer le risque avec des données socio-économiques issues du recensement de la population et des expositions environnementales comme la pollution de l'air, les niveaux de bruit et la proximité aux industries polluantes, au trafic automobile ou aux espaces verts.

Deux volets principaux composent cette thèse. Le volet analyse de données détaille la mise au point d'une procédure de création d'indices socio-économiques multidimensionnels et la conception d'un package du logiciel R l'implémentant, puis la création d'un indice de multi-expositions environnementales. Dans cette partie, on utilise des techniques d'analyse de données pour synthétiser l'information afin de fournir des indicateurs composites utilisables directement par les décideurs publics ou dans le cadre d'études épidémiologiques. Le second volet concerne les modèles bayésiens et explique le modèle « BYM ». Celui-ci permet de prendre en compte les aspects spatiaux des données et est mis en œuvre pour estimer le risque de mortalité infantile.

Dans les deux cas, les méthodes sont présentées et différents résultats de leur utilisation dans le contexte ci-dessus exposés. On montre notamment l'intérêt de la procédure de création d'indices socio-économiques et de multi-expositions, ainsi que l'existence d'inégalités sociales de mortalité infantile dans les agglomérations étudiées.

Mots clés : analyse de données, modèles bayésiens, inégalités sociales de santé, expositions environnementales

Data analysis technics and bayesian models applied to the contexte of social health inequalities and environmental exposures

The purpose of this thesis is to improve the knowledge about and apply data mining techniques and some Bayesian model in the field of social and environmental health inequalities. On the neighborhood scale on the Paris, Marseille, Lyon and Lille metropolitan areas, the health event studied is infant mortality. We try to explain its risk with socio-economic data retreived from the national census and environmental exposures such as air pollution, noise, proximity to traffic, green spaces and industries.

The thesis is composed of two parts. The data mining part details the development of a procedure of creation of multidimensional socio-economic indices and of an R package that implements it, followed by the creation of a cumulative exposure index. In this part, data mining technics are used to synthesize information and provide composite indicators amenable for direct usage by stakeholders or in the framework of epidemiological studies. The second part is about Bayesian models. It explains the "BYM" model. This model allows to take into account the spatial dimension of the data when estimating mortality risks.

In both cases, the methods are exposed and several results of their usage in the above-mentioned context are presented. We also show the value of the socio-economic index procedure, as well as the existence of social inequalities of infant mortality in the studied metropolitan areas.

Keywords : data analysis, Bayesian models, social health inequalities, environmental exposures

Laboratoires

Institut Élie Cartan de Nancy
Université de Lorraine
B.P. 70239
54506 Vandoeuvre-lès-Nancy Cedex

École des Hautes Études en Santé Publique, Rennes-Sorbonne Paris Cité
Avenue du Professeur Léon-Bernard
CS 74312
35043 Rennes Cedex

Table des matières

Table des figures

Liste des tableaux

Liste des principales abréviations

AASQA	Association agréée de surveillance de la qualité de l'air
ACP	Analyse en composantes principales
AFCM	Analyse factorielle des correspondances multiples
AFM	Analyse factorielle multiple
AIC	*Akaike information criterion*
CAH	Classification ascendante hiérarchique
CNIL	Commission nationale de l'informatique et des libertés
CRAN	*Comprehensive R archive network*
CSTB	Centre scientifique et technique du bâtiment
CépiDC	Centre d'épidémiologie sur les causes médicales de décès
DIC	*Deviance information criterion*
EHESP	École des hautes études en santé publique
HCSP	Haut conseil de la santé publique
IGN	Institut national de l'information géographique et forestière
Inpes	Institut national de prévention et d'éducation pour la santé
INSEE	Institut national de la statistique et des études économiques
Inserm	Institut national de la santé et de la recherche médicale
MCMC	*Markov chain Monte Carlo*
OMS	Organisation mondiale de la santé
SIDA	Syndrome de l'immunodéficience acquise
SMR	*Standardised mortality ratio*
VIH	Virus de l'immunodéficience humaine

Introduction générale

Introduction générale

Les mathématiques appliquées, et notamment les statistiques, fournissent sans cesse de nouveaux modèles, techniques et méthodes qu'un grand nombre de domaines (scientifiques, économiques, industriels) s'approprient ensuite afin de répondre à leurs problématiques propres. Cependant, le délai entre la mise au point de ces méthodes et leur adoption « courante » (voire leur adaptation) dans un domaine où elles peuvent s'avérer utiles est parfois très long et peut se compter en dizaines d'années.

C'est ainsi parfois le cas en santé publique. La santé publique est un domaine par nature multidisciplinaire qui implique de nombreux champs de recherche, ce qui engendre une multiplication des difficultés liées à chacun de ces champs. Ainsi, dans le cadre d'études s'intéressant aux inégalités sociales, répondre à des questions comme la définition précise des notions de statut socio-économique ou de population défavorisée n'est pas forcément simple pour un chercheur en santé publique et peut nécessiter par exemple d'impliquer des experts en sciences sociales. De même, les études impliquant l'environnement peuvent introduire des difficultés liées à la mesure et à l'estimation fiable de différentes expositions et nuisances, nécessitant l'intervention d'experts de ces domaines.

Les difficultés statistiques apparaissent souvent à la fin d'une longue chaîne d'autres complications et il n'est pas rare de voir des études qui, après avoir établi d'importants et complexes protocoles pour répondre à ces dernières, se contentent d'analyses statistiques relativement simples. C'est probablement l'une des raisons qui explique que les domaines de l'épidémiologie sociale et environnementale n'utilisent pas encore de manière régulière des techniques statistiques « récentes » permettant pourtant de prendre en compte de multiples problématiques qu'ils rencontrent.

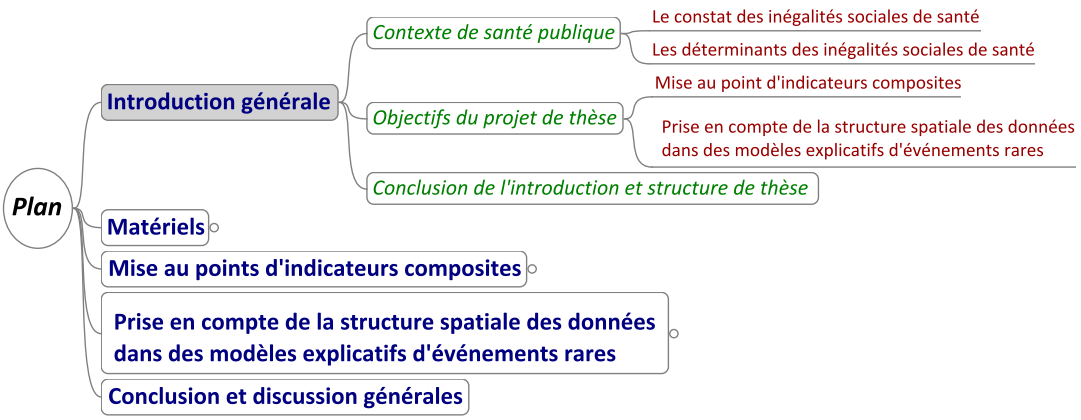
> C'est à la frontière entre les mathématiques appliquées et la santé publique que cette thèse s'est déroulée, visant à améliorer les connaissances de ce second domaine vis-à-vis des différentes méthodologies qui peuvent y être appliquées. Et en particulier afin de déterminer si l'apport de nouvelles techniques statistiques compense les inconvénients d'une mise en place plus complexe par rapport aux techniques « traditionnellement » utilisées.

Dans ce manuscrit :

- Avant de nous intéresser aux méthodes, nous commencerons par introduire le contexte de santé publique dans lequel s'inscrira cette thèse afin de mieux en saisir les problématiques et nous exposerons les objectifs de ce travail.
- Nous expliciterons dans une première partie les zones d'étude et l'échelle d'analyse, et définirons les données sanitaires, socio-économiques et environnementales choisies

et analysées.

- En seconde partie, nous détaillerons les méthodes étudiées, expliciterons les choix méthodologiques qui ont été faits au regard de considérations statistiques et des objectifs de santé publique visés, et présenterons différents travaux réalisés pour répondre au premier objectif : la mise au point d'indicateurs composites.
- En troisième partie, nous ferons de même avec le second objectif : la prise en compte de la structure spatiale de données dans les modèles explicatifs du risque d'événements rares.
- Enfin, nous conclurons le document.



Chapitre 1

Contexte de Santé Publique

Dans cette première section, nous présenterons le contexte de santé publique dans lequel s'inscrit cette thèse. Il sera tout d'abord question du constat des inégalités sociales de santé, notamment des inégalités de mortalité infantile, puis des déterminants sociaux de la santé et du rôle que peuvent y jouer certaines expositions environnementales.

1.1 Le constat des inégalités sociales de santé

1.1.1 Inégalités sociales dans le monde, en Europe et en France

De nos jours, le constat des inégalités sociales de santé est bien établi scientifiquement et l'ensemble des décideurs politiques et des acteurs de santé publique a pris conscience de ce phénomène. Il est ainsi amplement documenté à toutes les échelles que les populations les plus défavorisées sont davantage touchées par de nombreuses pathologies que les populations plus aisées, comme par exemple les maladies cardiovasculaires [1–5], les maladies respiratoires [2, 5–7], le diabète [2, 5], l'obésité [8], certains cancers [2, 5, 9, 10], la santé mentale [11, 12], les issues de grossesse défavorables et la mortalité infantile [13–16], etc. Les populations défavorisées ont également une espérance de vie et une espérance de vie en bonne santé plus courtes [17–19].

Ces inégalités sont observées aussi bien entre pays qu'au sein même de ceux-ci et à des échelles diverses [17, 20] : citons les inégalités de santé entre les pays industrialisés et les pays en voie de développement. Mais des inégalités de santé existent également à l'intérieur des pays suivant différents facteurs socio-économiques (revenu, éducation, ...) et à de fines échelles géographiques. Ceci impose une prise en compte de ces inégalités aussi bien dans les instances internationales que par les acteurs nationaux et locaux.

À l'échelle internationale, selon l'Organisation Mondiale de la Santé (OMS) l'espérance de vie à la naissance des hommes en 2011 allait de 46-48 ans pour des pays comme le Sierra Leone ou la République Centrafricaine, jusqu'à 80 ans (et plus) pour de nombreux pays dont l'Australie, la Suède, Israël, l'Islande ou le Qatar [21]. Parallèlement, les produits intérieurs bruts par habitant de ces pays, pour la même année, étaient respectivement de 501 $ et 488 $ pour les deux premiers ; et de 62 002 $, 57 071 $, 31 252 $, 44 120 $ et 90 524 $ pour les suivants [22].

Dans son rapport sur les statistiques du monde de 2007 [23], l'OMS montrait que l'espérance de vie en bonne santé à la naissance des femmes en 2002 pouvait aller de 35-38 ans pour l'Afghanistan, l'Angola ou le Mali jusqu'à 75 ans et plus pour des pays comme le Japon, l'Espagne, la Suisse, Monaco ou Andorre. Là encore, on peut noter en parallèle les taux d'alphabétisation (population de 15 ans et plus sachant lire et écrire) dans ces pays qui étaient respectivement de 28,1% (2000), 70,1% (2010) et 27,7% (2009) pour les trois premiers pays ; et de plus de 97% pour les suivants [24].

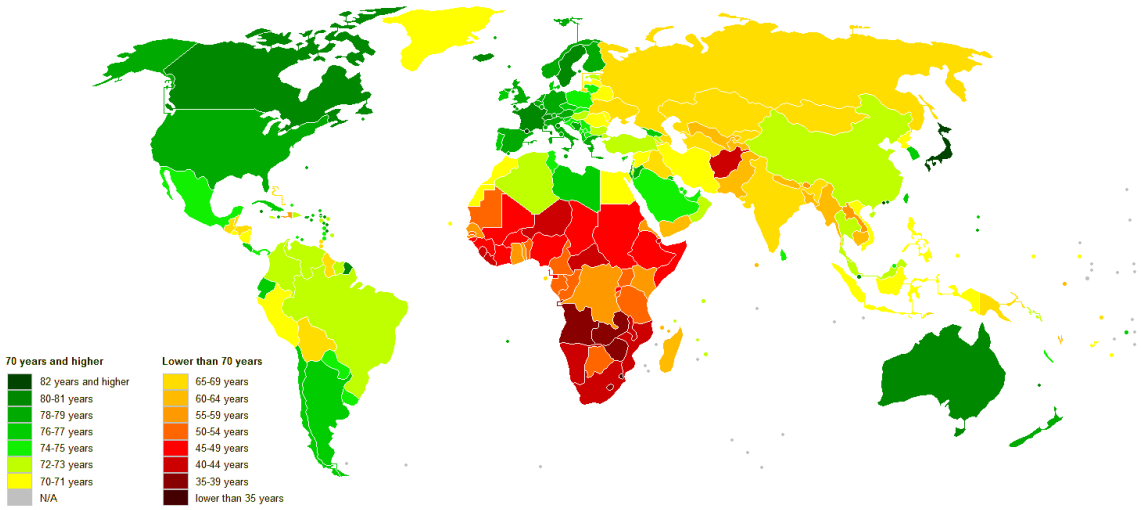


FIGURE 1.1 – Espérances de vie à la naissance en 2007 dans le monde.

Source : Wikimedia Commons & CIA World Factbook

Toujours selon l'OMS, la probabilité de décès entre 15 et 60 ans pour les deux sexes en 2011 allait de 55% en Lesotho à 5% en Islande, au Koweït, en Suisse ...[21] La mortalité d'enfants de moins de 5 ans était, elle, de 109 décès pour 1 000 naissances vivantes en 2011 en Afrique subsaharienne tandis qu'elle était de 7 pour 1 000 naissances vivantes dans les régions développées [25]. Les inégalités de santé et les inégalités socio-économiques observées entre les pays développés et ceux en voie de développement sont flagrantes (voir figure 1.1).

Les inégalités sociales de santé existent aussi au sein même des pays. Aux États-Unis un rapport du National Center for Health Statistics de 2013 [26] montre que le taux de mortalité (ajusté par l'âge) toutes causes en 2010 était de 741,8 pour 100 000 personnes chez les blancs contre 898,2 chez les noirs. En particulier, ces derniers présentaient des taux de mortalité de 224,9 pour 100 000 pour les maladies cardiaques (contre 176,9 pour les blancs) ou de 11,6 pour 100 000 dans le cas des maladies liées au VIH (contre 1,4 pour les blancs). En Nouvelle-Zélande, les statistiques sur la période 2010-12 [27] montrent que l'espérance de vie des Maoris était d'environ 7 ans inférieure à celle des non Maoris. Dans ces deux exemples, ce constat est associé au fait que ces populations (afro-américaine dans le premier cas et Maori dans le second) sont également en moyenne plus défavorisées, voire socialement exclues, illustrant donc l'aspect social de ces inégalités.

En France, il est fréquemment fait mention (y compris dans les médias) de l'existence d'importantes inégalités d'espérance de vie entre les différentes catégories socio-professionnelles (voir figure 1.2). Ainsi, sur la période 2000-2008 l'espérance de vie à 35 ans des hommes cadres supérieurs (47 ans) était de six ans supérieure à celle des ouvriers [19, 28]. Mais la durée de vie sans incapacité diffère également suivant les catégories socio-professionnelles, constituant ainsi une « double peine » : les cadres supérieurs peuvent ainsi espérer vivre 72% (34 ans) de leurs 47 années d'espérance de vie à 35 ans sans aucune incapacité, tandis que les ouvriers de 35 ans passeront eux en moyenne 24 (59%) de leurs 41 années de vie restante sans aucune incapacité [18].

Ces inégalités de santé sont également rapportées pour de nombreuses pathologies. L'étude ObEpi [29] a ainsi montré qu'en 2012 la prévalence de l'obésité dans la population française dépendait fortement, parmi d'autres facteurs, du niveau d'éducation, sa prévalence allant de 24,5% pour une éducation de niveau primaire à 14,3% pour un niveau d'éducation Baccalauréat et jusqu'à 7,3% dans le cas d'un diplôme d'enseignement supérieur de 3e cycle. D'autres études ont pu montrer le lien entre le niveau d'éducation, le revenu ou la catégorie socio-professionnelle et le tabagisme [30], la santé bucco-dentaire [31], les maladies professionnelles [32], etc.

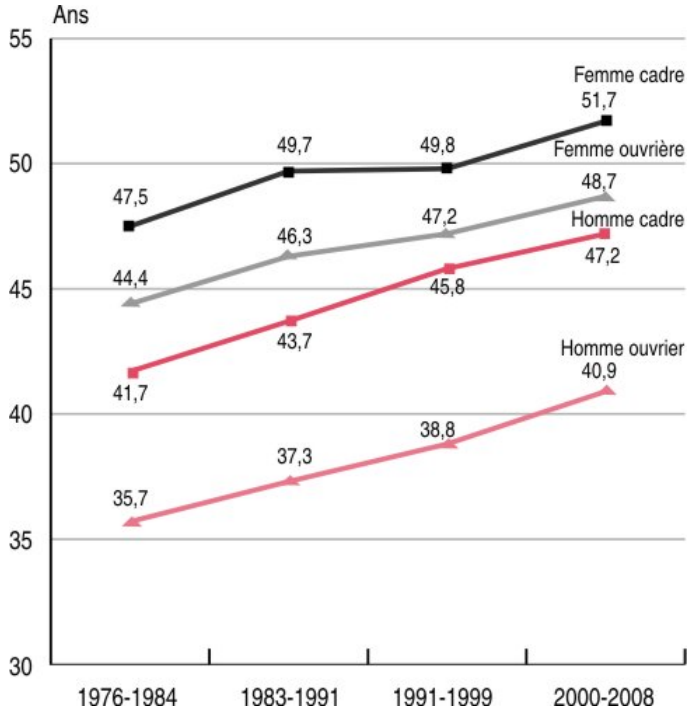


FIGURE 1.2 – Espérance de vie à 35 ans par sexe pour les cadres et les ouvriers.

Source : INSEE, Échantillon démographique permanent

Le constat des inégalités sociales de santé existe aussi à des échelles géographiques fines, entre quartiers d'une même agglomération par exemple. Ainsi, les travaux de Barcelò *et al.* [33] ont révélé l'existence d'associations significatives entre différentes causes de mortalité (liées notamment à la consommation d'alcool et de tabac chez les hommes, et liées à l'alimentation pour les femmes) et le statut socio-économique à l'échelle des quartiers

(*census tracts*) dans l'agglomération de Barcelone. Dans une revue systématique de la littérature suivie d'une méta-analyse, Meijer *et al.* [10] ont montré qu'il existait bien un lien entre le statut socio-économique des quartiers et la mortalité, avec des associations plus fortes pour les hommes et les jeunes. En France, à Strasbourg, une association entre l'infarctus du myocarde et le statut socio-économique des quartiers de résidence a été démontrée, le risque augmentant graduellement à mesure que le statut socio-économique diminuait [34].

Il faut en effet bien comprendre que, malgré les exemples donnés ci-dessus opposant des « extrêmes » de l'échelle du statut socio-économique, il existe bel et bien un gradient tout au long de celle-ci, qui est observé dans tous les pays [20].

Ce contexte est maintenant bien connu et fait de l'étude des inégalités sociales l'un des thèmes majeurs de la recherche en santé publique actuelle. L'accroissement des connaissances sur les déterminants de ces inégalités sociales de santé et la mise en place de moyens pour lutter contre celles-ci sont désormais fortement appelés aussi bien par les scientifiques eux-mêmes [17, 20, 35] que par les instances politiques internationales et nationales. Ainsi, parmi les huit objectifs Millénaire pour le Développement [36] établis par l'OMS, les trois premiers (éradiquer l'extrême pauvreté et la faim, assurer l'éducation primaire pour tous, promouvoir l'égalité des sexes et l'autonomisation des femmes) ont un lien direct avec les inégalités sociales de santé. Plus précisément, en mai 2012 les états membres de l'OMS ont adopté la résolution WHA65.8 [37] qui réaffirmait leur « *détermination à agir sur les déterminants sociaux de la santé* » et priaient notamment le directeur général de « *prendre dûment en considérations les déterminants sociaux de la santé dans l'évaluation des besoins sanitaires mondiaux* » et de « *continuer à faire mieux comprendre et souligner l'importance d'intégrer la question des déterminants sociaux de la santé dans les prochaines réunions des Nations Unies et autres réunions de haut niveau en rapport avec la santé et/ou le développement social* ».

Cette volonté de lutter contre les inégalités sociales se retrouve également au sein de l'Union Européenne à travers plusieurs déclarations, textes et commissions comme le second programme d'action communautaire dans le champ de la santé 2008-2013 [38], qui fait de la réduction des inégalités de santé l'un de ses objectifs, ou encore comme la déclaration intitulée « Solidarité en matière de santé : réduction des inégalités de santé dans l'Union européenne » [39] qui définit les grandes questions à traiter et actions à mener en la matière. On observe également une telle volonté au sein du bureau européen de l'OMS, comme cela a pu être déclaré en 2010 à Parme [40].

En France, la question des inégalités sociales de santé est abordée par de nombreuses institutions et est présente dans de nombreux plans d'actions, mais aussi dans la loi. Ainsi, la loi n°2004-806 du 9 août 2004 relative à la politique de santé publique [41] inclut parmi ses dispositions « *la réduction des inégalités de santé, par la promotion de la santé, par le développement de l'accès aux soins et aux diagnostics sur l'ensemble du territoire* » mais aussi et surtout un « *principe de réduction des inégalités : principe selon lequel la définition des objectifs et l'élaboration des plans stratégiques doivent systématiquement prendre en compte les groupes les plus vulnérables en raison de leur exposition à des déterminants spécifiques de la fréquence et/ou de la gravité du problème visé, y compris les déterminants liés à des spécificités géographiques* ».

Parmi les différents plans d'actions, on peut citer le Programme National Nutrition Santé 2011-2015 [42] dont le premier axe est de « *réduire par des actions spécifiques les inégalités sociales de santé dans le champ de la nutrition au sein d'actions générales de prévention* ». C'est également une thématique forte du Plan Cancer 2009-2013 [43] dont l'axe Observation vise notamment à soutenir « *l'épidémiologie sociale des cancers, qui devra permettre de mieux connaître et caractériser les inégalités sociales et territoriales d'incidence, de mortalité, d'exposition aux risques et d'accès à la prévention, au dépistage et aux soins* » et dont l'axe Prévention Dépistage inclut des mesures visant à lutter contre les inégalités (par exemple la mesure 14 « Lutter contre les inégalités d'accès et de recours au dépistage »).

Différents organismes comme l'Institut national de prévention et d'éducation pour la santé (Inpes) ou le Haut Conseil de la Santé Publique (HCSP) évoquent par ailleurs régulièrement la thématique des inégalités sociales de santé dans leurs avis et rapports [44–50].

Afin de bien mesurer les liens entre les inégalités sociales et la santé des populations, il convient cependant d'avoir des indicateurs pour mesurer cette dernière.

1.1.2 Indicateurs de l'état de santé des populations

La question des indicateurs de santé est extrêmement vaste et dépasse le cadre de cette thèse. En effet, elle englobe aussi bien des questions d'évaluation de l'état de santé des populations vis-à-vis de différentes pathologies, d'indicateurs de structure des systèmes de soin, de qualité des traitements, ou encore de performance des hôpitaux. Les domaines d'utilisation de ces indicateurs sont également très étendus, puisqu'ils vont de l'épidémiologie à l'évaluation de politiques publiques, en passant par l'économie de la santé. Ainsi, Murray *et al.* [51] listaient en 2000 les différentes utilisations possibles de ces mesures de synthèses :

1. Comparer la santé d'une population avec celle d'une autre
2. Surveiller les changements dans la santé d'une population donnée
3. Identifier et quantifier les inégalités de santé globale au sein des populations
4. Fournir une attention appropriée et équilibrée aux effets de santé non fatals sur la santé globale de la population
5. Informer les débats sur les priorités de planification des services de santé
6. Informer les débats sur les priorités pour la recherche et le développement
7. Améliorer le parcours d'enseignement professionnel en santé publique
8. Analyser les bénéfices d'interventions de santé pour des analyses coût-efficacité.

Ici, nous chercherons à expliciter uniquement certains des indicateurs de santé publique capables de donner une bonne appréciation de l'état de santé général des populations et de la qualité générale des systèmes de santé d'un point de vue épidémiologique, c'est-à-dire essentiellement les points 1 à 4. Les indicateurs couramment utilisés dans ce cas sont l'espérance de vie, les indicateurs de mortalité prématurée, ainsi que différents taux de morbidité ou de mortalité [52].

L'espérance de vie est un indicateur bien connu du grand public. L'indicateur de ce type le plus souvent utilisé est l'espérance de vie à la naissance, que l'on peut définir comme le nombre moyen d'années qu'un nouveau-né appartenant à une génération donnée (l'ensemble des personnes nées la même année) pourra vivre sous l'hypothèse que la structure de la mortalité du moment demeure stable. Différentes autres « variantes » existent, comme l'espérance de vie pour une année donnée (espérance de vie à 35 ans, par exemple) ou encore les différents types d'espérance de vie prenant en compte la qualité de vie (espérance de vie en bonne santé ou espérance de vie corrigée de l'incapacité). Ces indicateurs permettent donc de mesurer de manière générale la durée moyenne de vie d'une population (ou de sous-groupes de celle-ci), tout en répondant à la question (dans le cas des indicateurs ajustés par la qualité de vie) de savoir si l'on vit non seulement plus longtemps, mais aussi plus longtemps en bonne santé.

Les indicateurs de mortalité prématurée ou d'années potentielles de vie perdues ont pour but d'estimer le nombre d'années qu'un sujet ne vivra pas. Il s'agit de mesurer la différence entre l'âge réel de décès et un âge de référence (souvent 65 ans), afin d'estimer la mortalité liée à certaines causes. Les accidents, les suicides ou certaines pathologies spécifiques font ainsi perdre davantage d'années de vie potentielles.

Les différents taux de mortalité fournissent également des informations importantes. Le taux de mortalité brut est fréquemment utilisé mais il peut être également calculé pour des tranches d'âges particulières (mortalité infantile, mortalité des moins de 5 ans) ou des causes spécifiques, aussi bien transmissibles (VIH/SIDA, Malaria) que chroniques (maladies cardiovasculaires, cancers, diabète, ...). Dans la mesure où la seule mortalité n'est parfois pas suffisante, des indicateurs de morbidité sont parfois également utilisés (par exemple les taux d'incidences des maladies citées ci-dessus)

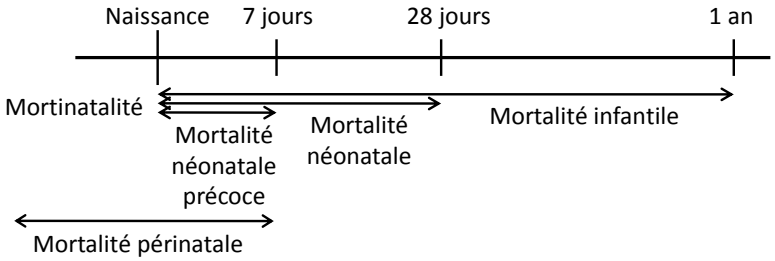


FIGURE 1.3 – Définitions des différentes mortalités en début de vie.

Parmi l'ensemble de ces indicateurs, le taux de mortalité infantile (nombre de décès d'enfants de moins d'un an pour mille naissances vivantes, voir figure 1.3) est considéré comme l'un des meilleurs pour refléter l'état de santé général des populations et la qualité des systèmes de santé [53]. En effet, les liens entre les causes de mortalité infantile et d'autres facteurs capables d'influencer la santé de populations entières (comme le développement économique, les conditions de vies en général, le bien-être social, la qualité de l'environnement, l'accès aux soins) font que les mesures prises pour réduire ce taux de mortalité peuvent rejaillir sur l'ensemble de la population. Cet indicateur est par ailleurs simple à obtenir (relativement à d'autres mesures plus élaborées) et permet, dans la mesure où il

est sensible aussi bien aux changements structurels qu'aux épidémies, d'avoir des informations rapides sur les changements dans la santé des populations (contrairement à d'autres indicateurs plus élaborés) [53].

L'importance de cet indicateur est par ailleurs également soulignée à différents niveaux : la réduction de la mortalité infantile (et celle des enfants de moins de 5 ans, dite « juvénile ») est le quatrième objectif du Millénaire pour le Développement [36] dont la liste est définie par l'OMS. En France, la loi du 9 août 2004 relative à la santé publique [41] inclut le taux de mortalité infantile parmi les indicateurs d'importance car il « *mesure non seulement la santé infantile, mais reflète aussi l'état de santé d'une population ainsi que l'efficacité des soins préventifs et l'attention accordée à la santé de la mère et de l'enfant. Cet indicateur rend compte en outre de facteurs sociaux plus larges tels que le niveau de scolarité des mères ou leur situation socio-économique* » et il « *reflète à la fois les conséquences des conditions de vie et celles des soins préventifs et curatifs accordés aux mères et aux enfants* ».

La mortalité néonatale (décès d'enfants de moins de 28 jours) représente une part importante des décès inclus dans la mortalité infantile. En effet, parmi l'ensemble des décès d'enfants de moins d'un an, plus de 50% (voire 60%) se produisent durant le premier mois de vie, période particulièrement « critique » [16, 54, 55].

L'étude des inégalités de mortalité infantile et néonatale peut ainsi permettre d'améliorer la lutte contre la réduction des inégalités sociales de santé en général.

1.1.3 Les inégalités sociales de mortalité infantile et néonatale

De la même manière que dans le cadre général présenté dans la section 1.1.1, on retrouve pour la mortalité infantile d'importantes inégalités à de nombreuses échelles, aussi bien entre pays qu'entre quartiers.

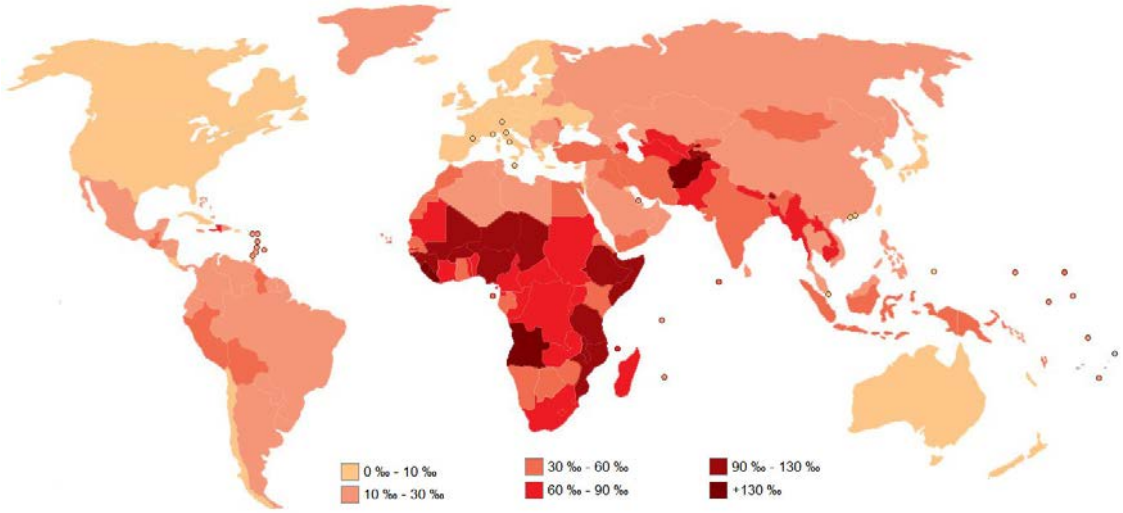


FIGURE 1.4 – Taux de mortalité infantile dans le monde.

Source : Wikimedia Commons & CIA World Factbook

À l'échelle internationale existent déjà de très importantes variations entre pays, en particulier entre les pays développés et ceux en voie de développement (voir figure 1.4). Ainsi, selon l'estimation 2013 faite par le CIA World Factbook [24] les taux de mortalité infantile en Afghanistan, au Mali et en Somalie étaient respectivement de 119, 106 et 102‰ naissances vivantes tandis que la Suède, Singapour et le Japon avaient des taux inférieurs à 3‰ (respectivement 2,73, 2,59 et 2,17) et que Monaco passait même sous la barre des 2‰ avec seulement 1,81 décès d'enfants de moins d'un an pour mille naissances vivantes. Dans ces mêmes pays, les taux de mortalité néonatale en 2011 étaient eux respectivement de 36, 49 et 50‰ naissances vivantes pour les trois premiers ; et de 1 ou 2‰ pour les suivants [22].

On constate également des inégalités entre les pays d'Europe, quoique moins marquées. Ainsi, le dernier rapport européen de santé périnatale [55] donne pour 2010 des taux de mortalité infantile (respectivement néonatale) allant de 9,8‰ (resp. 5,5‰) naissances vivantes en Roumanie à 2,3‰ (resp 1,5 et 1,2‰) naissances vivantes pour la Finlande et l'Islande.

Au sein même des pays, de nombreuses études ont également mis en évidence les disparités de mortalité infantile et néonatale. Le rapport européen de santé périnatale cité ci-dessus détaille ainsi les différents taux de mortalité infantile (respectivement néonatale) présents au Royaume-Uni : tandis que l'Angleterre, l'Ecosse et le Pays de Galles ont ainsi des taux très voisins de 3,7‰ (resp. 2,5‰), l'Irlande du Nord a elle un taux de 5,4‰ (resp. 3,8‰). En Belgique, ce taux est respectivement de 3,1, 3,3 et 4,8‰ (resp. 2,1‰, 2,3‰ et 2,7‰) en Wallonie, en Flandres et à Bruxelles.

Différentes études ont par ailleurs confirmé les liens entre la mortalité infantile et différentes composantes du statut socio-économique. Aux États-Unis, la dernière version d'un rapport établi annuellement par le ministère de la santé [54] indique qu'en 2007 les taux de mortalité infantile étaient de 9,17‰ pour les mères non-mariées et 5,16‰ pour les mères mariées, le statut marital étant alors considéré comme un marqueur de la présence ou de l'absence de ressources financières et sociales. La même étude montre un gradient de mortalité infantile en fonction du niveau d'éducation de la mère, décroissant de 7,8‰ pour les mères avec moins qu'un diplôme de lycée jusqu'à 3,8‰ pour les mères ayant une licence ou plus.

En Europe, une étude au Pays de Galles [56] a montré que le risque combiné de mortinatalité et de mortalité infantile augmentait de 53% dans les quartiers les plus défavorisés par rapport aux quartiers les plus favorisés. En Norvège, Arntzen *et al.* [15] ont montré sur une étude de toutes les naissances entre 1967 et 1998 que si le risque de mortalité infantile s'était considérablement réduit sur la période étudiée, il demeurait inversement proportionnel au niveau d'éducation des parents.

En France, même si le taux de mortalité infantile a été divisé par trois entre 1980 et 2010 [57] il a désormais tendance à décroître moins rapidement que dans les autres pays d'Europe, voire à stagner (voir figure 1.5). Ainsi, le rapport européen de santé périnatale cité précédemment [55] indique un taux de mortalité infantile (resp. néonatale) en France de 3,5‰ (resp. 2,3‰) en 2010, tandis que sa version précédente [58] donnait un taux de 3,9‰ (resp. 2,6‰) en 2004. Parallèlement, entre 1999 et 2009 la France est passée du

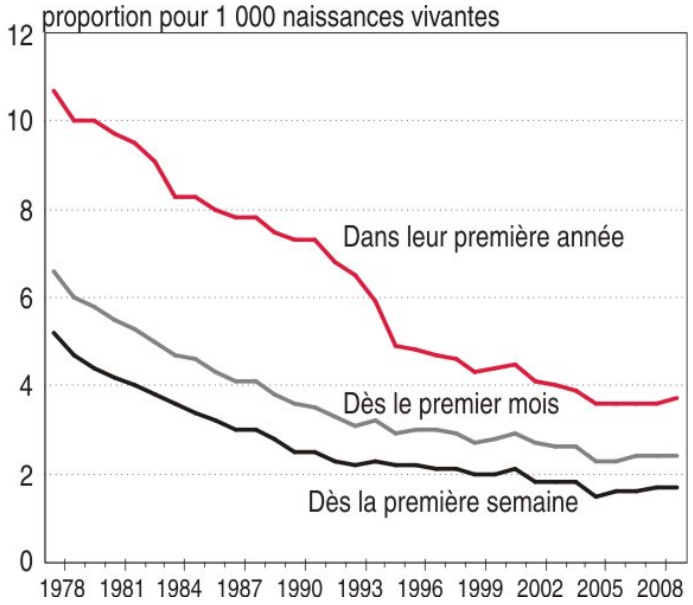


FIGURE 1.5 – Évolution des taux de mortalité infantile, néonatale et néonatale précoce en France entre 1978 et 2008.

Source : INSEE, statistiques de l'état civil

7e au 20e rang européen pour la mortalité infantile (de 8e ex-aequo au 17e rang pour la mortalité néonatale), bien qu'il soit tout de même à noter que les taux de nombreux pays européens sont tellement proches qu'une légère variation du taux français peut entraîner d'importants changements dans le classement.

On observe cependant d'importantes disparités sur le territoire français, par exemple entre régions. Ainsi le taux moyen de mortalité infantile entre 2007 et 2009 allait de 2,8‰ en Corse à 4,8‰ en Alsace, et même jusqu'à 8,8‰ pour l'ensemble des Départements d'Outre-Mer. Ces différences peuvent en partie être expliquées par les différents aspects socio-économiques de ces régions (certains de ceux-ci seront illustrés dans la section 3.2.1).

Et de la même manière que dans les autres pays, on trouve des liens forts entre la mortalité infantile et les facteurs socio-économiques. Ainsi, un rapport de l'Institut national de la statistique et des études économiques (INSEE) de 2011 [57, 59] se basant sur l'Échantillon Démographique Permanent a montré une association significative entre la mortalité infantile et la catégorie sociale de la mère et du père. Ce même rapport note que, bien que les inégalités sociales se soient fortement resserrées avec le temps, le taux de mortalité infantile en 2009 des femmes cadres étant de 2,6‰ , celui des enfants de femmes employées de 3‰ et celui des enfants de femmes ouvrières de 3,5‰.

1.2 Les déterminants des inégalités sociales de santé

1.2.1 Les déterminants connus de la littérature

De nombreux déterminants sociaux de la santé sont désormais bien connus dans la littérature [60]. Ainsi, différents domaines peuvent avoir des effets directs ou indirects

sur la santé :

- la stabilité économique : statut d'emploi, revenu, stabilité de logement ...
- l'éducation,
- le contexte familial et/ou communautaire : structure familiale, cohésion sociale, exclusion ...
- le système de soins : accès aux services de soins primaire, préventif, clinique ...
- l'environnement de logement et de quartier : qualité de logement, accès aux magasins de première nécessité, criminalité et violence, conditions environnementales ...

Dans le cas particulier de la mortalité infantile, les déterminants sociaux ont une influence majeure [55]. Des associations ont été notamment montrées avec le niveau d'éducation de la mère [16, 54, 55, 61] ou des deux parents [15], la catégorie sociale des parents [55, 57], le statut marital de la mère [54], le revenu du foyer ou encore le statut socio-économique du quartier de résidence [56, 61]. Par ailleurs, la mortalité infantile est également influencée par d'autres facteurs (obésité, diabète, tabagisme des parents, ...) qui sont eux-mêmes fortement associés aux facteurs sociaux [8, 62, 63].

Cependant, malgré ces différents déterminants sociaux déjà connus, il reste une part inexpliquée dans les inégalités sociales de santé. Depuis plusieurs années, les expositions environnementales sont suspectées comme pouvant expliquer en partie cette zone d'ombre [64, 65].

1.2.2 Les expositions environnementales

Débutée en Amérique du Nord dans les années 1980 et plus récemment étendue en Europe, la cause de la « justice environnementale » vise à étudier les liens entre les expositions environnementales et le statut socio-économique des populations. À l'origine, ces études se concentraient généralement sur les liens entre la résidence de minorités ethniques, souvent à revenus faibles, et leur proximité par rapport à certaines industries, axes de trafic, décharges mais elles incluent d'autres types de pollutions (comme les différents polluants de l'air). La littérature documente désormais de mieux en mieux les liens entre indicateurs socio-économiques et expositions environnementales, comme les décharges toxiques, la pollution de l'air intérieur et extérieur, la pollution de l'eau ou encore le bruit [64, 66].

On distingue essentiellement deux mécanismes qui pourraient expliquer les liens complexes entre l'environnement, le statut socio-économique et la santé : le différentiel d'exposition et le différentiel de vulnérabilité[65].

Le différentiel d'exposition suppose que les déterminants sociaux affectent le lieu de vie des individus et peuvent par conséquent participer à ce que certaines populations (*a priori* défavorisées) soient plus exposées à des conditions environnementales nuisibles à la santé par rapport à d'autres groupes (*a priori* plus favorisés) [65]. Ceci peut être dû aussi bien au fait que les populations défavorisées viennent habiter dans des zones plus exposées (pour des raisons économiques, par exemple) que réciproquement par le fait que l'implantation de nouvelles installations polluantes se fasse plus facilement dans des quartiers de faible niveau socio-économique. Par ailleurs, en complément du contexte de vie, les déterminants sociaux

peuvent également affecter directement l'exposition aux nuisances environnementales via le type d'emploi, l'éducation ou la qualité du logement par exemple.

Ce mécanisme est déjà relativement bien étudié : de manière générale les populations les plus défavorisées vivent plus proches de nombreuses sources de pollution ou dans des lieux où les niveaux d'exposition sont plus importants [67–70]. Ainsi, Braubach et Fairburn [69] ont publié une revue de la littérature montrant les liens entre un faible statut socio-économique et l'exposition à la pollution de l'air intérieur, au bruit, à la proximité aux industries polluantes et au manque d'espaces verts. Briggs *et al.* [68] ont quant à eux montré des liens entre le revenu, l'éducation, l'emploi et des variables de pollution de l'air extérieur (particules, dioxyde d'azote, dioxyde de soufre), la proximité au trafic, aux industries et aux décharges. La même étude montre cependant que, dans certains cas ou pour certaines expositions (le radon par exemple), ce ne sont pas forcément les plus défavorisés qui sont les plus exposés. D'autres exemples de ce type existent : ainsi à Rome, il a été montré que les populations les plus favorisées étaient plus exposées à la pollution de l'air [71] ; à Strasbourg, ce sont plutôt les classes intermédiaires qui subissent davantage cette pollution [72].

Le différentiel de vulnérabilité (ou de susceptibilité) suppose quant à lui que certaines populations subiront, pour une même exposition, davantage de problèmes de santé que d'autres. Cette vulnérabilité peut être liée à des caractéristiques « intrinsèques » (âge, sexe, génétique, ethnicité ...) mais surtout à des caractéristiques « extrinsèques » incluant en particulier les déterminants sociaux. Ainsi, les populations les plus défavorisées auraient davantage de problèmes de santé liés à cette exposition que les populations plus aisées [65]. Ceci peut-être lié au fait que ces populations défavorisées ont déjà, en partie à cause de ces mêmes déterminants sociaux, une santé plus dégradée qui les rend plus sensibles aux effets sanitaires des expositions. Cela peut également être lié au fait que les déterminants sociaux affectent l'accès à certains produits de première nécessité ou aux soins [73].

Ce mécanisme est moins facile à mettre en évidence mais a également déjà été étudié. Ainsi, O'Neill *et al.* [74] ont montré dans plusieurs villes des États-Unis que la mortalité due aux fortes chaleurs était plus importante chez les personnes moins éduquées que chez celles avec une éducation niveau lycée ou supérieur, ainsi que chez les noirs par rapport aux blancs, confirmant d'autres résultats de ce type. Des différences d'effet de l'exposition à la pollution de l'air sur la santé suivant le statut socio-économique individuel ou du quartier ont également été relevées dans plusieurs études [71, 75].

En agissant de manière indépendante ou combinée, ces deux mécanismes pourraient expliquer la part des inégalités sociales de santé encore inconnue. Néanmoins, il demeure des incertitudes sur ce lien pour de nombreux types d'expositions et de nombreuses agglomérations. De la même manière, les parts respectives du différentiel d'exposition et du différentiel de vulnérabilités ne sont pas forcément bien déterminées.

Chapitre 2

Objectifs du projet de thèse

Dans ce contexte, l'objectif de cette thèse est d'améliorer les connaissances sur les différentes méthodologies applicables aux problématiques des études d'inégalités socio-territoriales et environnementales de santé, afin de **déterminer si l'apport de nouvelles techniques statistiques compense les inconvénients d'une mise en place plus complexe par rapport aux techniques plus couramment employées dans le domaine**. Cette thèse s'intègre dans les perspectives du projet Equit'Area [76] qui a pour but général d'améliorer les connaissances et la compréhension des relations entre santé, environnement et caractéristiques socio-économiques et qui constituera le cadre d'application.

Deux volets principaux sont développés, constituant la structure de ce manuscrit :

- D'une part, la **mise au point d'indicateurs composites** synthétisant l'information contenue dans d'importants jeux de données, illustrée ici par la construction d'indices de statut socio-économique et d'indices de multi-expositions environnementales. La réalisation de cet objectif conduira à appliquer des méthodes d'analyse de données et à développer des procédures utilisant celles-ci.
- D'autre part, la **prise en compte de la structure spatiale des données dans des modèles explicatifs du risque d'événements rares**, avec comme illustration l'exploration des liens entre la mortalité infantile ou néonatale, le statut socio-économique et les expositions environnementales. Cet objectif nous mènera à étudier et appliquer les modèles bayésiens.

2.1 Mise au point d'indicateurs composites

A l'heure actuelle, de très nombreuses bases de données sont disponibles en France, notamment grâce à l'essor de l'informatique et d'Internet, mais également dans le cadre du mouvement qui veut rendre de plus en plus accessible au public les informations issues de la statistique publique, notamment dans le cadre de l'environnement [77]. Cependant, ces bases de données sont très dispersées tant leurs sources sont nombreuses. Les différents ministères, institutions ou centres de recherches proposent des bases de données issues aussi bien d'études spécifiques que réalisées « en routine » avec une mise à jour régulière [59, 78–80].

La recherche fait parfois partie des utilisations prévues initialement pour ces bases de données, mais ce n'est souvent pas le cas pour celles existant depuis longtemps et qui ont fréquemment été conçues pour d'autres fins comme par exemple la surveillance, la gestion ou la vérification du suivi de la réglementation sanitaire. L'ouverture de beaucoup de ces bases de données aux chercheurs, voire au grand public, est malgré tout une énorme source d'informations qui peuvent venir grandement enrichir la recherche sur de nombreux sujets.

Cependant, une fois ces bases de données obtenues et réunies (ce qui n'est pas trivial étant donné l'hétérogénéité de leurs formats, échelles géographique et temporelle ...), cette profusion d'informations peut vite devenir problématique pour le chercheur. Ainsi, les résultats du recensement comportent par exemple plus de 1800 indicateurs socio-économiques. Utiliser l'ensemble de ces indicateurs devient alors difficile voire impossible, que ce soit par comparaison avec le nombre d'unités statistiques de l'étude (qui peut être inférieur au nombre de variables) ou à cause des corrélations importantes entre les indicateurs (qu'on doit donc éviter d'utiliser simultanément dans un modèle classique).

Il est alors parfois compliqué de choisir quelles données utiliser et la manière de traiter celles-ci. La sélection et/ou la synthèse des informations contenues dans ces bases de données importantes deviennent alors particulièrement nécessaires et requièrent l'emploi de techniques adaptées. Or, dans de nombreux domaines d'application ces techniques sont peu connues ou maîtrisées et ne sont donc pas utilisées.

> Dans ce contexte, la nécessité de l'utilisation des techniques d'analyse de données, qui sont précisément conçues pour gérer de grandes bases de données et en extraire les informations les plus pertinentes, paraît évidente. C'est pourquoi le premier volet de cette thèse ayant pour but de mettre au point des indicateurs composites à partir d'importants jeux de données de sources diverses a conduit à étudier, choisir et appliquer ce type de techniques.

2.2 Prise en compte de la structure spatiale des données dans des modèles explicatifs du risque d'événements rares

Actuellement, de nombreux projets de recherche en épidémiologie sociale ou environnementale sont de design « écologique » de type géographique, c'est-à-dire qu'ils ont pour unités statistiques des zones géographiques plutôt que des individus. Cela résulte du développement de plus en plus large et varié de grandes bases de données statistiques par les institutions publiques et parapubliques et, à l'opposé, aux difficultés et aux coûts auxquels sont confrontées les études fondées sur la collecte de données individuelles.

Or une dépendance spatiale peut exister entre ces zones, deux unités voisines ayant généralement des caractéristiques plus proches que deux zones très distantes l'une de l'autre. L'hypothèse d'indépendance des observations faite dans de nombreux modèles statistiques classiquement utilisés dans ce domaine n'est donc pas respectée dans ce cas.

Par ailleurs, dans le cas de l'étude du risque d'événements rares, ce qui est le cas de la mortalité infantile (et encore plus de la mortalité néonatale) en France, la moindre fluctuation dans le nombre de cas observés de l'événement (que cette fluctuation soit aléatoire ou non) a une influence importante sur le risque de celui-ci, l'estimation de ce dernier pouvant alors avoir une forte variabilité et imprécision.

Ces deux particularités nécessitent donc d'adapter les techniques habituellement utilisées en épidémiologie ou d'employer des techniques différentes. Au sein du projet dans lequel s'inscrit cette thèse, ceci s'est traduit par l'utilisation de modèles statistiques ou géostatistiques prenant en compte les composantes spatiales sous différentes formes, comme des modèles additifs généralisés, des modèles de balayage géographique ou certains types de modèles bayésiens. Ce sont ces derniers qui ont été plus particulièrement étudiés dans notre cas.

Conclusion de l'introduction et structure de thèse

Pour atteindre l'objectif cité, cette thèse va s'atteler à la mise au point d'indicateurs statistiques composites et prendre en compte la structure spatiale des données dans des modèles explicatifs de risques d'événements rares. Ceci impliquera d'appliquer et d'améliorer les connaissances concernant les techniques d'analyse de données et certains modèles bayésiens dans un domaine d'étude peu familier avec celles-ci. En guise d'illustration, ces différentes techniques seront appliquées dans le cadre de l'étude des inégalités sociales et environnementales de santé, avec comme objet la mortalité infantile et néonatale sur les agglomérations Marseille, Lyon et Lille et sur la ville de Paris durant la période 2000-2009.

Dans ce contexte, la suite de ce manuscrit sera structurée de la façon suivante.
La partie I explicitera le **matériel d'étude** en définissant d'abord l'unité d'analyse et en détaillant les zones d'étude sélectionnées. Ensuite, chacune des familles de données (sanitaires, socio-économiques, environnementales) sera abordée en détaillant leurs sources, leur construction et en explicitant la définition des différentes variables choisies et employées.

La partie II détaillera la **mise au point d'indicateurs composites**, avec les détails du développement d'une procédure de création d'indices socio-économiques et la conception d'un package du logiciel R l'implémentant, puis avec la création d'un indice de multi-expositions environnementales, et enfin avec quelques perspectives sur l'analyse de données multi-niveaux.

La partie III concernera la **prise en compte de données spatiales dans l'étude du risque d'événements rares** et plus particulièrement les modèles bayésiens, dont le modèle « BYM » et quelques perspectives sur les modèles « ZIP ».

Première partie

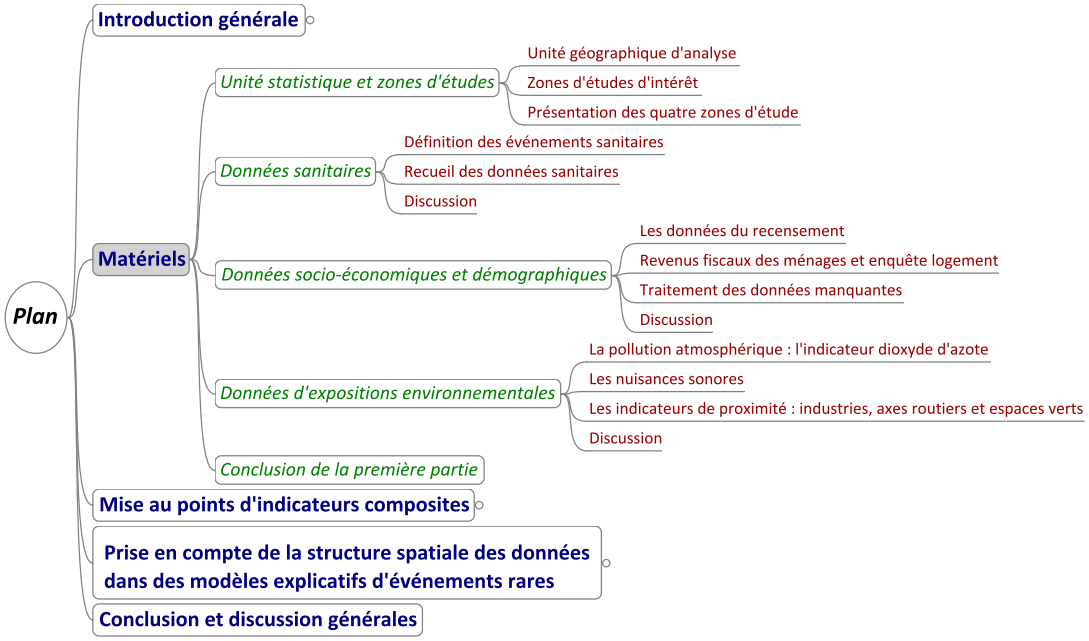
Matériels

Matériels

Toute étude statistique nécessite de définir la mesure de variables sur des unités statistiques à partir des données.

Dans cette première partie, nous présenterons les différentes sources de données disponibles au sein du projet Equit'Area et plus particulièrement les données exploitées dans le cadre de cette thèse et nous expliciterons les raisons pour lesquelles elles ont été choisies.

Pour cela, nous définirons tout d'abord les unités statistiques et géographiques retenues ainsi que les zones d'études d'intérêt. Ensuite, les différentes variables mesurées ou estimées sur ces zones seront précisées en suivant la répartition suivante : (1) les données sanitaires, (2) les données socio-économiques et démographiques, et (3) les données environnementales.



Chapitre 3

Unité statistique et zones d'étude

Dans ce chapitre, nous définirons l'unité géographique et statistique d'analyse retenue, puis après avoir justifié le choix des zones d'étude, nous présenterons ces dernières plus en détail.

3.1 Unité géographique d'analyse

3.1.1 IRIS : Îlots Regroupés pour l'Information Statistique

L'IRIS, acronyme de « Îlots Regroupés pour l'Information Statistique », est une unité géographique infra-communale créée par l'INSEE à l'occasion du recensement de la population française de 1999. Cette unité est approximativement équivalente aux *census block groups* et aux *census tracts* anglo-saxons, fréquemment utilisés dans les études du même type [5, 33, 68, 81–84].

Établi par l'INSEE en collaboration avec la CNIL et les collectivités locales, un IRIS regroupe en moyenne 2 000 habitants dans ce qui est désormais souvent considéré comme l'échelon de base pour les analyses statistiques géographiques. C'est, en tout cas, **la plus fine unité pour laquelle l'INSEE fournit des données statistiques détaillées**.

La détermination des contours des IRIS n'est pas réalisée uniquement sur des critères statistiques et géographiques. Bien qu'ils doivent « *respecter des critères géographiques et démographiques et avoir des contours identifiables sans ambiguïté et stables dans le temps* », ils sont aussi construits de manière à représenter ce que l'INSEE désigne comme des « quartiers vécus » [85].

C'est ainsi que l'INSEE décrit que « *le découpage en IRIS est donc avant tout la remontée de la vision de l'espace communal par les communes elles-mêmes qui étaient au mieux à même de fournir le découpage le plus utile. Fruit de cette collaboration, les IRIS possèdent, individuellement, une identité forte. Espaces géographiques connexes sans ruptures internes, ils correspondent plus à des espaces facilement identifiables par les habitants de la commune et possèdent un nom généralement évocateur* » [85].

À partir de ces critères, toutes les communes de plus de 10 000 habitants et une bonne partie de celles de plus de 5 000 habitants ont vu leur territoire partitionné en IRIS.

La France (DOM inclus) a ainsi été divisée en environ 16 000 IRIS, auxquels il convient de rajouter toutes les communes non découpées (environ 34 800), assimilées à des IRIS uniques, ce qui permet de recouvrir ainsi tout le territoire français d'un maillage de l'ordre de 50 000 unités géographiques.

3.1.2 Définition des différents types d'IRIS

La diversité de l'organisation du territoire impose néanmoins de distinguer trois catégories d'IRIS définies par l'INSEE [85] :

- Les **IRIS d'habitation** (92% des IRIS en 2008) regroupent entre 1 800 et 5 000 habitants, sont homogènes vis-à-vis du type d'habitat et ont des contours basés sur les grandes coupures du tissu urbain. Il s'agit donc aussi bien de quartiers d'habitation d'un centre-ville que de zones pavillonnaires, et leur superficie varie en conséquence.
- Les **IRIS d'activité** regroupent plus d'un millier de salariés, qui doivent en outre être deux fois plus nombreux que la population résidente. Ce sont donc, par exemple, les zones d'activité ou les quartiers d'affaires.
- Enfin, les **IRIS divers** recouvrent une grande surface peu peuplée, comme les parcs de loisirs, les forêts ou encore les zones portuaires.

Malgré la part relativement faible d'IRIS d'activité et divers, cette classification en plusieurs types est particulièrement importante à prendre en compte car les profils de ces IRIS vis-à-vis de certaines variables, particulièrement celles considérant le nombre d'habitants, sont très marqués.

> Cette distinction en trois catégories d'IRIS aux caractéristiques distinctes nous conduira à effectuer différents choix méthodologiques (considérer certains types comme illustratifs, comme on le verra en section 7.3) ou à guider notre interprétation des résultats (comme par exemple section 8.3).

3.1.3 Choix des IRIS

Par conception, l'étude qui a servi d'application à cette thèse est une étude « écologique » de type géographique, c'est-à-dire où l'unité statistique n'est pas l'individu mais une aire géographique (souvent définie suivant des contours administratifs). Le choix de cette aire, qui constituera l'unité statistique et géographique est déterminant [86]. Ce choix est fondé autant sur des considérations statistiques que sur des aspects liés aux applications proprement dites, chaque échelle géographique, de la région à l'IRIS, possèdant ses propres avantages et inconvénients.

Disponibilité des données

L'accessibilité des données fait partie des critères permettant de guider le choix. Ainsi, à l'heure actuelle, la plus fine échelle pour laquelle il est possible d'obtenir de nombreuses données est l'IRIS, défini dans le paragraphe précédent. On sait par ailleurs qu'il existe au sein même des communes une variabilité entre les IRIS la composant aussi bien du point de vue des caractéristiques socio-économiques qu'environnementales (ainsi qu'on le remarquera par l'analyse descriptive de certaines de ces données en section 3.2 mais aussi à l'aide d'analyses multi-niveaux au chapitre 9).

Biais écologique

L'un des problèmes lié aux études écologiques est le « biais écologique » [87]. Celui-ci, qui peut aussi être assimilé à un biais d'agrégation, est un problème d'interprétation qui peut apparaître lorsque l'on tente de faire des déductions au niveau individuel à partir d'inférences faites sur les groupes auxquels les individus appartiennent. Bien qu'il ne soit pas possible de le supprimer entièrement (sauf en revenant à l'échelle individuelle), réduire la taille des unités géographiques analysées permet, en regroupant moins d'individus et en réduisant donc leur hétérogénéité, de réduire les problèmes liés à ce biais. Choisir une échelle spatiale d'analyse fine permet donc d'obtenir une plus grande précision des résultats. Il devient alors possible d'identifier plus clairement les zones « à risque » ou ayant des profils particuliers, et par conséquent d'aider à établir des actions plus ciblées.

L'un des avantages de cette échelle d'analyse est que, bien qu'il s'agisse d'une unité dont le découpage est en partie administratif (comme la commune par exemple), la construction des IRIS prend en compte la taille de la population et l'homogénéité de celle-ci dans ses caractéristiques démographiques et socio-économiques (au contraire de la commune), ce qui devrait permettre de minimiser d'autant le biais écologique.

Cette homogénéité des caractéristiques est certainement moins bien respectée lorsqu'il s'agit d'IRIS faiblement peuplés ; en effet, l'IRIS s'apparente alors entièrement à la commune et ses contours deviennent totalement dépendants de ceux définis administrativement. Cette difficulté pourrait être particulièrement cruciale dans des études intégrant des zones géographiques à la fois rurales et urbaines puisque la définition de l'IRIS serait variable.

> Pour ces raisons et parce que peu d'autres études en France ont travaillé à ce niveau géographique dans ce domaine, ce projet travaille au niveau de l'IRIS. L'utilisation d'aires géographiques comme unités statistiques constitue néanmoins une particularité qu'il nous a fallu prendre en compte dans l'ensemble de ce travail (et qui justifie par ailleurs l'un des points principaux de cette thèse). Nous détaillerons plusieurs fois dans la suite de quelle manière la structure spatiale a influencé ce travail, que ce soit dans la construction de variables environnementales (chapitre 6), l'étude des caractéristiques socio-économiques (chapitre 9) ou dans les modèles explicatifs de risque (chapitre 10).

3.2 Zones d'étude d'intérêt

3.2.1 Le choix de plusieurs zones d'étude

Deux sources de variabilité spatiale sont intégrées dans ce travail : la première est liée à la diversité des caractéristiques des IRIS composant une zone d'étude ; la seconde correspond aux contrastes pouvant exister entre zones d'étude, ce qui nous a conduit délibérément à en sélectionner plusieurs. Plus précisément, différentes agglomérations (zones urbaines) ont été retenues afin d'augmenter la variabilité des facteurs de risques qui nous intéressent dans notre application (c'est-à-dire les caractéristiques socio-économiques et environnementales) et ainsi mieux expliquer les différences de variation de l'événement sanitaire considéré.

L'objectif est alors d'étudier les disparités socio-spatiales qui existent entre les agglomérations et en leur sein. *D'emblée, les zones rurales ont été exclues de ce travail.* En effet,

bien qu'il paraisse important d'étudier les inégalités sociales et environnementales de santé dans un contexte rural, les déterminants qui sont en jeu dans ces deux contextes, rural et urbain, sont très différents et ne peuvent être combinés au sein d'une même analyse. La problématique des inégalités spatiales en zones rurales et la manière d'adapter le travail présenté dans ce manuscrit à celle-ci font néanmoins partie des réflexions qui sont actuellement menées, ainsi qu'on le verra par la suite dans la section 7.7.

3.2.2 Choix des agglomérations

Différentes méthodes peuvent être mises en œuvre pour la sélection des agglomérations d'intérêt. L'une de ces méthodes consisterait à dresser une typologie (par classification par exemple) des agglomérations françaises à partir d'informations caractérisant leur niveau socio-économique et démographique, et leur niveau d'exposition environnementale (pour ne retenir que celles particulièrement approfondies dans ce travail). Sélectionner aléatoirement ou par choix raisonné une agglomération parmi chaque profil obtenu permettrait de maximiser la diversité des agglomérations de l'étude.

Cependant, cette méthode requiert d'avoir dès le début une base de données suffisamment riche pour effectuer une classification ; ce qui n'était pas notre cas au début de la thèse. De plus, il fallait garantir la sélection d'agglomérations de taille importante permettant la collecte de données sanitaires en nombre suffisant pour conduire une analyse de puissance statistique raisonnable.

> C'est pour cette raison que les agglomérations ont été retenues suivant un choix raisonné parmi celles dont les contrastes sanitaires et environnementaux étaient documentés et qui se trouvent par ailleurs être les quatre plus grandes agglomérations françaises. Il s'agit de la ville de Paris et de sa Petite Couronne (départements des Hauts-de-Seine, de la Seine-Saint-Denis et du Val-de-Marne), de l'unité urbaine de Marseille-Aix-en-Provence, de l'agglomération du Grand Lyon et de l'agglomération de Lille Métropole.

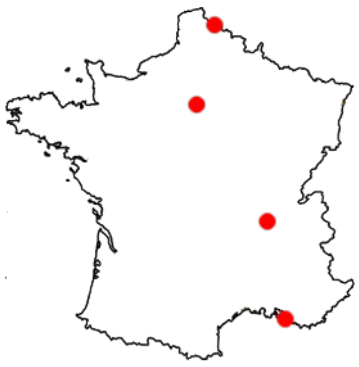


FIGURE 3.1 – Localisation des quatre agglomérations retenues

Ces agglomérations présentent des caractéristiques différentes dans l'organisation de leur territoire et de leur urbanisme, dans leur situation sanitaire, dans leurs particularités démographiques et socio-économiques ou encore dans la diversité des contextes environnementaux (voir table 3.1). Elles regroupent environ 10,4 millions d'habitants en 2008, soit un peu moins de 17% de la population française.

La diversité dans leurs caractéristiques permettra de comparer leurs profils et de chercher à déterminer les aspects qui leur sont communs (ces aspects pourraient être supposés existants pour d'autres agglomérations) et leurs particularités propres. Les connaissances accrues spécifiques à une agglomération pourront aider les pouvoirs publics locaux à mettre en œuvre des solutions. Par ailleurs, la proportion de la population prise en compte via ces quatre agglomérations est très importante, ce qui implique par conséquent un plus grand nombre d'événements sanitaires inclus dans l'étude.

TABLE 3.1 – Caractéristiques démographiques et socio-économiques des quatre agglomérations

	Paris	**Marseille**	**Lyon**	**Lille**
Population (2008-2009) (habitants)	6 578 258	1 563 036	1 281 971	1 105 080
Nombre de communes	124	38	58	85
Nombre d'IRIS	2 749	630	510	506
Superficie (km^2)	760	1 290	527	611
Chômage (%)[a]	10,56	12,65	9,93	12,09
Ouvriers (%)[a]	8,85	12,24	12,34	16,67
Familles monoparentales (%)[a]	8,78	10,55	8,39	9,81
Sans diplôme (%)[a]	13,17	16,61	13,15	16,03
HLM (%)[a]	13,21	5,81	11,65	15,08
Revenu annuel médian par UC (€)[a]	21 241	17 131	18 849	16 368

Source : INSEE

Pour des raisons de place, seul le nom des villes principales des agglomérations a été indiqué.

[a]En 2006. Médiane des valeurs des IRIS de la commune.

3.3 Présentation des quatre zones d'étude

3.3.1 Paris - Petite Couronne

Paris, en tant qu'unité administrative, est la commune la plus peuplée de France (2 211 297 habitants en 2008 pour une densité de 20 169 hab/km^2). Elle est aussi l'unique « commune-département » du pays, au centre d'une agglomération qui figure parmi les plus peuplées d'Europe et qui s'étend sur presque toute la région Île-de-France. Dès le Xe siècle, la position de Paris en fait l'une des principales villes de France, dont l'influence en Europe et dans le monde ne cessera quasiment pas de grandir tout au long du millénaire. La richesse historique et culturelle de la ville, son influence sur les sciences, l'éducation, les arts et les divertissements ainsi que son poids économique et politique en Europe, font toujours de Paris l'une des villes majeures du monde occidental. La commune est divisée en 20 arrondissements, comptant chacun entre 14 et 96 IRIS et conduisant Paris à compter 992 IRIS.

Les frontières administratives de la capitale n'ayant quasiment pas évolué depuis la fin du XIXe siècle, contrairement à certaines de ses homologues européennes, l'agglomération de Paris s'étend bien au-delà des limites de la commune de Paris elle-même et recouvre la totalité de la Petite Couronne (formée des trois départements limitrophes de Paris : les Hauts-de-Seine (92), la Seine-Saint-Denis (93) et le Val-de-Marne (94)) et une partie

de la Grande Couronne (composée des départements périphériques de l'Île-de-France : la Seine-et-Marne, les Yvelines, l'Essonne et le Val-d'Oise).

La Petite Couronne est une véritable extension de la ville, voire du centre-ville via certains quartiers comme La Défense qui est l'un des quartiers d'affaires majeurs en Europe. Sa population en 2008 était de 4 366 961 habitants répartis sur environ 655 km^2. Parmi les villes les plus peuplées (ayant parfois une densité de population supérieure à Paris elle-même), on peut citer Boulogne-Billancourt, Nanterre et Courbevoie pour les Hauts-de-Seine, Saint-Denis, Montreuil et Aubervilliers pour la Seine-Saint-Denis ou encore Créteil, Vitry-sur-Seine et Villejuif pour le Val-de-Marne. La Petite Couronne compte 123 communes divisées en 1 757 IRIS.

C'est, dans le cadre de l'étude Equit'Area, Paris et sa Petite Couronne qui ont été retenues, formant ainsi une zone d'environ 760 km^2 accueillant près de 6,6 millions d'habitants dans 124 communes (2 749 IRIS, voir figure 3.2).

FIGURE 3.2 – Communes et IRIS de Paris-Petite Couronne

3.3.2 Unité urbaine de Marseille - Aix-en-Provence

Marseille, préfecture de la région Provence-Alpes-Côte-D'azur, est l'une des plus anciennes villes de France. Située sur les rives de la Méditerranée, l'activité portuaire de la ville a contribué à façonner son histoire de l'Antiquité à nos jours. La ville est désormais l'un des principaux ports de France, de Méditerranée et d'Europe en termes de volume total de marchandises manipulées. En tant que ville portuaire, elle est également depuis

des siècles une ville de migrations venues de toute la Méditerranée. Importante ville industrielle jusqu'au XXe siècle, elle conserve un certain nombre d'industries sur son territoire malgré les crises qui ont pu toucher ce secteur. Partie intégrante d'une culture régionale riche et active, elle est par ailleurs la capitale européenne de la culture en 2013. Marseille est composée de 8 secteurs (de deux arrondissements chacun), divisés en 393 IRIS et sur lesquels vivaient, en 2009, 850 602 habitants.

FIGURE 3.3 – Communes et IRIS de l'unité urbaine d'Aix-Marseille

Autour de Marseille s'est construite la communauté urbaine de *Marseille Provence Métropole*, regroupant à sa création (en 2000) 17 communes voisines et un peu plus d'un million d'habitants. Cette structure politique ne reflète cependant pas la totalité de l'agglomération marseillaise. En effet, certaines municipalités de l'agglomération ont refusé de faire partie de *Marseille Provence Métropole* et ont constitué autour d'elles d'autres communautés de communes comme la *communauté d'agglomérations du pays d'Aix*, la *communauté d'agglomérations d'Aubagne et de l'Etoile* ou encore la *communauté d'agglomérations du pays de Martigues*. Ce grand nombre d'entités et les liens qu'elles entretiennent a malheureusement été la cause de plusieurs problèmes d'accessibilité des données pour cette agglomération.

L'unité urbaine de Marseille-Aix-en-Provence, que l'on peut rapprocher le plus de la notion d'agglomération, est donc bien plus étendue que Marseille seule ou Marseille Provence Métropole, et inclut également les autres communautés de communes précédemment citées, et en particulier Aix-en-Provence. C'est, en termes de population, la deuxième agglomération française après Paris avec 1 563 036 habitants répartis dans 38 communes et 630 IRIS (voir figure 3.3).

3.3.3 Grand Lyon

Lyon, préfecture de la région Rhône-Alpes, fut capitale des Gaules durant l'Antiquité romaine. Située au croisement du nord et du sud de l'Europe et au confluent de la Saône

et du Rhône, elle est proche aussi bien du nord de la France que de la Méditerranée, de l'Italie et du nord-est de l'Espagne. Cette position a contribué à faire de Lyon une ville économique prospère depuis le Moyen-Âge. Historiquement associée à l'industrie textile et pétrochimique, son industrie se recentre désormais davantage sur le domaine de la pharmacie et des biotechnologies. Lyon est également l'une des plus grandes villes universitaires de France, et son patrimoine architectural et culturel a conduit plusieurs de ses quartiers à être classés au patrimoine mondial de l'UNESCO. La commune est divisée en 9 arrondissements (185 IRIS) accueillant 479 803 habitants en 2009.

FIGURE 3.4 – Communes et IRIS du Grand Lyon

La communauté urbaine de Lyon, appelée plus généralement *Grand Lyon*, existe depuis la fin des années 1960. Elle englobe la majeure partie des banlieues lyonnaises à l'exception de quelques-unes des plus éloignées du centre et regroupe plus de 80% des habitants du département du Rhône dans moins de 15% de sa superficie. Un peu moins étendue que l'unité urbaine de Lyon, l'entité politique du *Grand Lyon* reflète néanmoins de manière correcte l'agglomération lyonnaise qui était, en 2009, la troisième agglomération de France en termes de population avec 1 281 971 habitants dans 58 communes et 510 IRIS (voir figure 3.4).

3.3.4 Lille Métropole

Lille est la préfecture de la région Nord-Pas-de-Calais. Située au croisement d'itinéraires terrestres et maritimes, elle est à la jonction de l'Europe du Nord et du Sud (Pays-Bas, Belgique, France) tout autant qu'entre l'Est et l'Ouest (Allemagne, Belgique, Luxembourg, Royaume-Uni). Cette position en a fait une ville marchande depuis plusieurs siècles, mais aussi un lieu contesté qui a souffert de nombreux conflits. Elle a également été une ville de manufactures, puis d'industries, notamment dans les domaines textile et mécanique, dont le déclin a plongé la ville dans la crise des années 1960 aux années 1990. La réhabilitation de certains quartiers ainsi que le développement de quartiers d'affaires et d'un pôle universitaire à partir de cette période ont néanmoins permis un certain renouveau. L'histoire riche et mouvementée de la ville lui a léguée un patrimoine, autant d'un point de vue architectural que culturel, qui lui a valu d'être capitale européenne de la culture en 2004. Lille comptait 226 827 habitants en 2009, répartis dans 95 IRIS.

FIGURE 3.5 – Communes et IRIS de Lille Métropole

Lille fait partie à l'heure actuelle d'une vaste mégalopole (ou plus exactement « conurbation ») comptant également Roubaix et Tourcoing, mais aussi plusieurs villes belges comme Tournai et Courtrai. Lille Métropole communauté urbaine retranscrit en partie, du côté français, cette conurbation en regroupant 85 communes dans la plus grande communauté urbaine de France en nombre de communes. Créée en 1996 (alors sous le nom de communauté urbaine de Lille), cette communauté urbaine est structurée autour du noyau Lille-Roubaix-Tourcoing et de leurs communes périphériques, en y ajoutant vers le sud un axe construit autour de Villeneuve d'Ascq et au nord un autre axe vers la frontière belge. L'agglomération lilloise est la quatrième plus peuplée de France et Lille Métropole comptait, en 2008, 1 105 080 habitants répartis dans 506 IRIS (voir figure 3.5).

3.4 Discussion

On a vu dans ce chapitre quels avaient été les choix d'échelle d'analyse et de zones d'étude effectués.

Ainsi, on a sélectionné parmi les différentes échelles disponibles (régionale, départementale, cantonale, communale, etc) l'échelle de l'IRIS. Celle-ci permet, par sa finesse et par sa construction, de réduire autant que possible le biais écologique pouvant apparaître si l'on cherche à interpréter au plan individuel des résultats obtenus à l'échelle contextuelle tout en demeurant une unité géographique pour laquelle de nombreuses données (principalement socio-économiques) sont disponibles. On verra cependant dans les chapitres 4 et 6 que cette disponibilité de données n'est pas forcément vraie dans tous les champs de la santé publique, et notamment que des difficultés ont été rencontrées pour estimer les données sanitaires et environnementales à cette échelle spatiale.

Une autre limite inhérente au choix de cette échelle d'analyse est qu'elle demeure liée aux délimitations administratives puisque un IRIS, unité infra-communale, ne peut pas, par définition, s'étendre sur plusieurs communes. Ceci implique donc que des « quartiers vécus » (tels qu'on les a présentés en section 3.1.1) qui s'étendraient sur deux communes voisines ne pourront pas être représentés par un même IRIS.

Dans ce travail, seules des zones urbaines ont été analysées ; les zones rurales ont été exclues. En effet, leur profil démographique et de nombreux facteurs associés (offre de soins, contexte économique, environnement ...) diffèrent profondément du contexte urbain. Si l'on a cherché à augmenter les contrastes vis-à-vis des caractéristiques socio-économiques et environnementales en incluant plusieurs agglomérations, intégrer également des zones rurales aurait risqué d'écraser toute nuance dans les analyses menées. A titre d'essai, on a ainsi pu remarquer que si l'on appliquait la procédure détaillée chapitre 7 à l'ensemble de la Bretagne, la ville de Rennes apparaissait entièrement comme « très défavorisée » (ce point sera détaillé davantage dans la discussion 7.7).

En conséquence, le choix s'est porté sur quatre agglomérations urbaines contrastées et comptant une importante population, ce qui permet ainsi d'obtenir à la fois un nombre d'IRIS et d'événements sanitaires suffisant.

Chapitre 4

Données sanitaires

Dans cette seconde section, nous définirons les événements sanitaires considérés dans l'étude et détaillerons le recueil de ceux-ci. Nous exposerons également différents calculs de taux de mortalité basés sur ces données.

4.1 Définition des événements sanitaires

Dans le cas de la mortalité entourant la naissance, il convient de fixer correctement le vocabulaire et de distinguer plusieurs termes : la mortinatalité, la mortalité néonatale, la mortalité périnatale et la mortalité infantile.

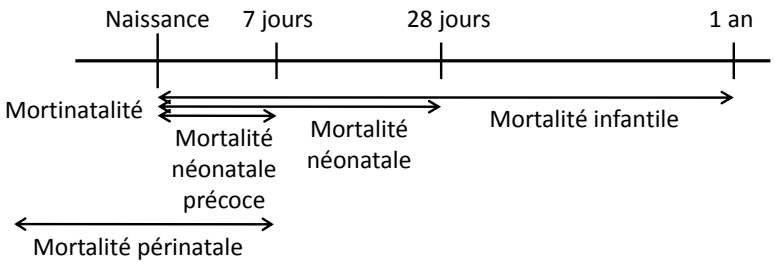


FIGURE 4.1 – Définitions des différentes mortalités en début de vie.

La mortinatalité correspond au nombre d'enfants mort-nés après au moins 24 semaines de grossesse, selon l'OMS. On appelle mortalité périnatale (exprimée pour 1 000 naissances totales) la réunion de la mortinatalité et de la mortalité néonatale précoce (voir ci-dessous).

Cependant, la définition de la mortinatalité et surtout le statut juridique des enfants mort-nés font l'objet de nombreuses évolutions, notamment d'un point de vue législatif et d'état civil, et reviennent périodiquement dans le débat public. La mortinatalité est définie différemment dans le monde. Au cours de la période de l'étude, il y a ainsi eu plusieurs modifications successives de la législation et de la jurisprudence en France [88] qui ont changé les différents critères déterminant si les enfants mort-nés pouvaient avoir un acte de naissance et de décès, un « acte d'enfant sans vie » ou encore faire l'objet d'une simple déclaration administrative.

Alors que la mortinatalité faisait partie des objets de départ de l'étude, la fluctuation importante de sa définition en France (fluctuation s'étendant par conséquent aux inscriptions dans les registres d'état civil) et les difficultés rencontrées dans le recueil des données sanitaires nous ont conduit à écarter cet événement et à ne considérer que la mortalité infantile et la mortalité néonatale. En effet, les définitions de ces événements sont stables et permettent donc d'avoir des données cohérentes sur l'ensemble de la période.

Comme énoncé dans l'introduction, la mortalité infantile est définie comme les décès d'enfants de moins d'un an. Le taux de mortalité infantile est exprimé pour 1 000 naissances vivantes et ne prend donc pas en compte les mort-nés. Ainsi que nous l'avons détaillé dans l'introduction, ce taux de mortalité est considéré comme un indicateur important de la qualité des soins obstétriques et pédiatriques en particulier, et de l'état de santé des populations en général.

La mortalité néonatale, exprimée pour 1 000 naissances vivantes, correspond aux décès situés entre la naissance et le 28e jour de vie. Elle est séparée en deux phases : la mortalité néonatale précoce, de la naissance au 7e jour de vie, et la mortalité néonatale tardive, du 7e au 28e jour. La mortalité néonatale est généralement considérée comme moins sujette aux inégalités sociales de santé que la mortalité post-néonatale, mais davantage soumise à l'influence du déroulement de la grossesse et des expositions environnementales [89, 90].

4.2 Recueil des données sanitaires

La constitution d'une base de données comprenant l'ensemble des cas de décès d'enfants de moins d'un an (ou de moins d'un mois) par IRIS sur les quatre zones d'étude du projet n'est pas triviale. Plusieurs semaines de travail de la part de plusieurs personnes (et personnellement deux semaines au début de cette thèse dans l'agglomération de Lyon) ont été nécessaires afin d'obtenir une telle base de données.

En effet, pour des raisons de confidentialité, les certificats de décès remplis par les médecins sont séparés en deux parties distinctes [91] :

- La première est utilisée pour la déclaration de décès à l'état civil et n'est pas anonyme (elle contient notamment l'adresse du domicile du décédé). C'est sur la base de cette partie que sera établi l'acte de décès qui contient des informations sur la date et le lieu de décès, sur l'identité de la personne décédée, celle de ses parents, ainsi que sur l'identité du déclarant. L'acte de décès est conservé dans les registres de décès de la commune de décès et est transmis à la commune de domicile de la personne décédée si celle-ci diffère (et est connue).
- La deuxième partie, anonyme, comporte des informations médicales sur les causes de la mort et sur les éventuels états morbides ou physiologiques associés ainsi que la date du décès et la commune où il est survenu. Cette partie est cachetée puis transmise par l'intermédiaire du médecin de santé publique de l'agence régionale de santé au Centre d'épidémiologie sur les causes médicales de décès (CépiDC) de l'Inserm qui, après analyse des diagnostics et codage, établit les statistiques nationales de mortalité.

Cependant, ces statistiques ne sont à l'heure actuelle disponibles en routine qu'à l'échelle communale (ou échelle supérieure). Une possibilité pour obtenir une estimation du nombre

de cas de mortalité infantile par IRIS aurait été de calculer ceux-ci au prorata des naissances dans l'IRIS à partir du nombre de cas dans la commune. Cette méthode aurait cependant conduit à supposer que le risque de mortalité était similaire entre IRIS d'une même commune et donc à gommer les contrastes entre ceux-ci, objet même de l'analyse.

Par conséquent, il a été choisi de retourner aux sources directes afin d'obtenir ces informations par IRIS, en se rendant dans les différentes mairies des communes composant les agglomérations pour examiner les registres de décès et en extraire les informations des actes de décès d'enfants de moins d'un an. Ces actes ont fourni des informations sur le sexe, les dates de naissance et de décès, l'adresse de résidence (des parents ou de la mère) et dans la moitié des cas la profession des parents.

En faisant l'hypothèse qu'en France la majorité des décès d'enfants se produit dans un centre de soins et sachant que les actes de décès sont établis dans la commune de décès, l'ordre de visite des communes a pu être optimisé en se rendant d'abord dans les communes possédant (ou ayant possédé sur la période d'étude) un centre de soins avec une unité d'obstétrique, puis celles possédant un centre de soins, puis les autres par taille de population. Ceci a permis d'établir, en accord avec la CNIL, une base de données des cas de mortalité infantile sur les différentes agglomérations.

En tant que partenaire du projet, le CépiDC a fourni l'ensemble des données de mortalité infantile à l'échelle de la commune, permettant d'avoir pour chacune le nombre de décès d'enfants de moins d'un an par année. Cette base de données complémentaire à celle qui a été constituée à l'IRIS a permis de vérifier l'exhaustivité des informations recueillies.

S'assurer que tous les cas ont pu être recueillis est particulièrement important dans ce contexte d'application. En effet, la mortalité infantile et néonatale étant des événements rares, l'absence d'un seul cas peut avoir des conséquences importantes sur l'estimation du taux de mortalité de l'IRIS et conduire à considérer qu'il présente ou non un excès de mortalité. À l'aide des données communales fournies par le CépiDC, il a donc été possible de vérifier que l'ensemble des cas de chaque commune avait pu être retrouvé dans les registres d'état civil. Il a ainsi été possible d'obtenir un taux d'exhaustivité des données de plus de 90% sur l'ensemble des agglomérations retenues (excepté pour l'agglomération d'Aix-Marseille où la collecte est encore en cours).

Une fois cette base de données à l'échelle individuelle constituée, afin de revenir à l'échelle de l'IRIS autant que pour respecter les consignes de confidentialité de la CNIL (notamment la suppression des adresses postales après le géocodage), chaque cas a été géocodé : l'adresse renseignée sur chaque acte de décès a ainsi été convertie en coordonnées géographiques puis affectée à un IRIS de manière semi-automatique grâce à des outils fournis par l'Insee et l'IGN (CAZU et contours IRIS) et exploités grâce à un système d'information géographique (ArcGIS).

Ce processus ne peut cependant pas être entièrement automatisé et souffre de certains problèmes de reconnaissance de la ville ou de la rue, liés à la diversité des notations possibles (utilisation d'abréviations entre autres), ce qui nécessite une longue préparation des données en amont. Certaines adresses ne peuvent par ailleurs pas être converties par le logiciel et nécessitent d'être attribuées à la main dans un IRIS.

4.3 Discussion

Ce chapitre a décrit comment les données sanitaires sur lesquelles ce travail a été appliqué ont été définies, recueillies et transformées en base de données exploitable.

Le choix de la mortalité infantile et néonatale implique différents aspects méthodologiques. S'agissant d'événements rares en France et l'unité géographique d'étude étant fine (ainsi qu'on l'a exposé au chapitre 3 précédent), le nombre d'événements par année est particulièrement faible. Pour cette raison, nous avons décidé que l'étude porterait sur les données cumulées sur les dix ans de la période 2000-2009. Ce choix permet ainsi d'augmenter le nombre de cas recueilli par IRIS et par conséquent d'augmenter la puissance statistique des analyses menées par la suite.

Néanmoins, même en considérant une période de dix ans, l'échelle d'étude est tellement fine que sur l'ensemble des zones d'étude le nombre maximum de cas dans un IRIS est de 7. L'événement demeure donc rare et impliquera d'utiliser des techniques statistiques adaptées. La prise en compte à la fois de ces événements rares et de la structure spatiale des données est ainsi le deuxième objectif de cette thèse et il sera présenté en partie III.

Chapitre 5

Données socio-économiques et démographiques

Dans cette section, nous présenterons les différentes données socio-économiques et démographiques utilisées dans l'étude, leurs sources et les traitements qui ont été réalisés.

5.1 Les données du recensement

La majorité des données socio-économiques analysées provient des recensements de la population française effectués par l'INSEE. En particulier, ce sont les données issues du recensement général de la population de 1999 et celles du recensement de la population de 2006. Ces données sont mises à disposition des chercheurs, et même souvent totalement publiques, via le site internet de l'INSEE ou de partenaire tels que le Réseau Quételet [78]. Différents stades de traitement et échelles sont disponibles, allant des données presque brutes (dénombrements) jusqu'à des indicateurs construits, pour des échelles allant de l'IRIS à la France entière.

5.1.1 1999, dernier recensement général de la population française

En 1999 a eu lieu le dernier recensement général de la population française, qui a concerné l'ensemble de la population au même moment via un comptage « traditionnel ». Ce type de recensement datait du début du XIXe siècle et s'est déroulé approximativement tous les cinq ans jusqu'en 1946. À partir de cette date, les intervalles entre deux recensements (jugés trop coûteux et complexes à mettre en œuvre) se sont peu à peu allongés pour aller de 6 à 9 ans, jusqu'aux derniers recensements généraux en 1990 et 1999. À ce moment, l'intervalle de temps entre deux recensements devint trop important par rapport à l'évolution démographique, ce qui a conduit à modifier la méthode mise en œuvre pour recenser la population française [92].

En effet, le 27 février 2002, la loi relative à la démocratie de proximité introduit une « rénovation du recensement » [93] et en définit les principes. À partir de 2004, le recensement de la population passe sur une base annuelle. Une nouvelle méthode est alors employée, impliquant une distinction entre les communes de plus de 10 000 habitants et celles de moins. Ces dernières sont regroupées en cinq groupes sur des critères statistiques leur assurant des poids démographiques similaires. Tous les ans, l'ensemble des communes de l'un des groupes est recensé exhaustivement. Par roulement, cela conduit à recenser en

cinq ans l'ensemble des communes françaises de moins de 10 000 habitants. Les communes plus grandes, quant à elles, voient leur population sondée tous les ans via un échantillon d'adresses tirées aléatoirement et représentant environ 8% de la population. En cinq ans, l'ensemble du territoire communal est pris en compte et environ 40% de la population auront été recensés [94].

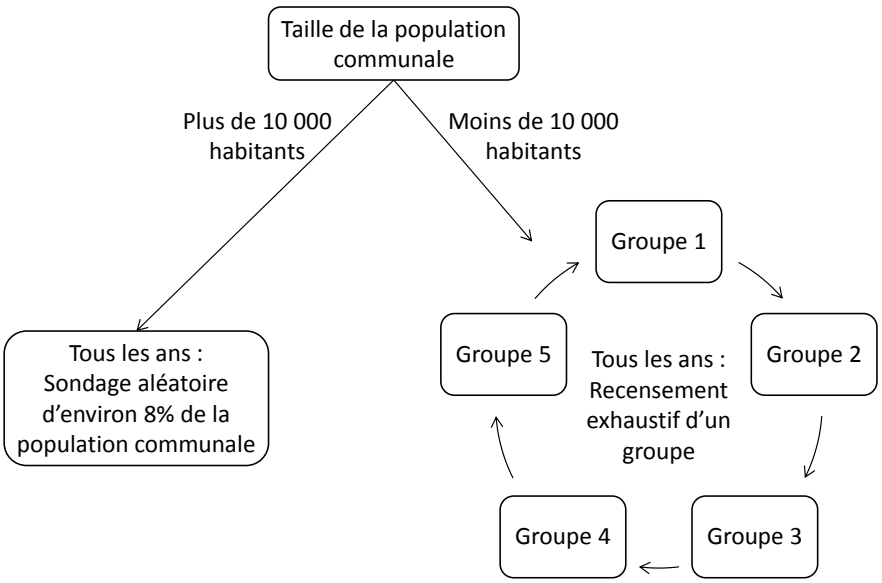


FIGURE 5.1 – Schéma d'organisation du recensement de la population après 2004.

Ce changement de pratique cherche ainsi à répondre à deux objectifs : tout d'abord mettre à disposition des données de façon plus fréquente et plus régulièrement actualisées, mais aussi réguler les coûts et charges liés à la mise en œuvre de ce recensement. Après une phase transitoire entre 2004 et 2008, c'est-à-dire durant le premier cycle quinquennal, qui a vu les résultats définitifs du « recensement millésimé 2006 » publiés fin 2008-début 2009, les populations légales et résultats statistiques sont désormais publiés tous les ans.

Au début de la thèse, les données disponibles les plus récentes étaient celles du recensement général de 1999, et c'est par conséquent celles-ci qui ont été utilisées en premier lieu. Par la suite, la mise à disposition des données du recensement 2006 a permis d'utiliser des données socio-économiques davantage en adéquation avec la période d'étude 2000-2009.

5.1.2 Présentation et justification des variables retenues

Les données de base fournies par l'INSEE pour le recensement 1999 sont regroupées en plusieurs fichiers thématiques (Population, Familles, Actifs, Logement, Formation, ...) donnant, pour chaque IRIS, les dénombrements d'un grand nombre de variables. On peut trouver, par exemple, le nombre de personnes de différentes tranches d'âges, niveaux de formation ou catégories socio-professionnelles. Plus de 2 000 indicateurs « bruts » sont ainsi présents dans les données de base issues du recensement.

Cependant, dans la mesure où il s'agit pour la grande majorité d'indicateurs donnant des valeurs absolues, ceux-ci ne nous intéressaient pas directement dans le cadre du projet.

Ils ont donc servi de point de départ à la construction d'une cinquantaine de nouvelles variables basées sur des taux, toujours à l'échelle de l'IRIS. On a choisi ici de ne construire que des variables quantitatives directement à partir des données brutes dans la mesure où il s'agissait de la manière la plus naturelle d'obtenir des indicateurs socio-économiques par IRIS à partir des données du recensement (ces données étant elles-mêmes quantitatives).

> Les variables initialement considérées dans l'analyse comme pouvant caractériser le mieux les contrastes socio-économiques et démographiques entre unités spatiales ont été sélectionnées suivant deux critères.

Tout d'abord, une revue de la littérature a permis d'identifier les variables classiquement introduites dans les études [82, 95–105]. Il s'agit par exemple des variables caractérisant le niveau d'éducation, la catégorie socio-professionnelle ou le revenu, qui sont les plus couramment utilisées.

Puis, d'autres variables ont été volontairement introduites car nous suspections qu'elles pouvaient être pertinentes dans le contexte territorial de notre étude ou pouvaient être considérées comme liées au concept habituel de statut socio-économique (et par conséquent donner également un éclairage sur le statut socio-économique contextuel). Ce sont par exemple des variables comme la proportion de personnes âgées de plus de 65 ans ou encore la proportion de personnes qui ont déménagé de leur commune depuis le dernier recensement.

De plus, l'objectif étant de constituer des indicateurs composites à l'aide de techniques appropriées, aucune limite de nombre de variables n'a été fixée à ce stade. Il s'agit ici d'une différence importante avec d'autres études s'intéressant aux inégalités sociales qui n'intègrent pas d'indicateurs composites et qui sont alors confrontées à de fortes contraintes de choix et de nombre de variables à intégrer dans un même modèle (dans ce cas, il est généralement préféré d'avoir un nombre restreint de variables soigneusement choisies suivant la littérature).

Toutes ces variables couvrent différents grands thèmes : structure familiale, type de foyer, statut migratoire, mobilité, emploi, éducation, logement, etc. Le but ici a été d'introduire un spectre de variables plus large que celui habituellement utilisé. En impliquant davantage de variables, cela ouvre la possibilité d'examiner l'utilité de prendre en compte certaines « approximations » mais aussi d'employer des techniques guidées par les données elles-mêmes.

C'est également pour cette raison que certaines variables sont volontairement « redondantes », au sens où elles sont différentes variantes d'une même notion. Par exemple, le chômage peut être décliné en chômage :

- total,
- de longue durée,
- chez les seniors,
- chez les juniors,
- chez les hommes,
- chez les femmes ...

En n'adoptant aucun *a priori*, ce sera également avec des techniques orientées données qu'il sera déterminé quelle est la « meilleure » variable pour représenter cette notion (par exemple, vaut-il mieux employer le taux de chômage total, celui chez les jeunes ou celui de longue durée ?). En effet, différentes analyses préliminaires ont pu montrer que ce n'était pas toujours l'indicateur le plus « intuitif » (par exemple le chômage total) qui représentait le mieux la notion sur certaines agglomérations (voir section 7.2.1).

5.1.3 Définitions des variables par thèmes et domaines

La liste des différents indicateurs construits à partir des données du recensement 1999 et leurs définitions est disponible tables 5.1 et 5.2.

TABLE 5.1 – Description des variables socio-économiques construites à partir des données du recensement 1999, par domaine

Famille et foyer	Moins de 25 ans dans la population totale Plus de 65 ans dans la population totale « Hors ménages » (élèves internes, militaires, détenus, résidents de maisons de retraite, foyers, hôpitaux, ...) dans la population totale Familles monoparentales parmi l'ensemble des familles ⋆ Ménages composés d'une personne seule parmi l'ensemble des ménages
Mobilité et immigration	Étrangers (personnes n'ayant pas la nationalité française) dans la population totale ⋆ Migrants (personnes vivant à l'étranger au dernier recensement, sans considération de nationalité) dans la population totale Immigrants étrangers (personnes n'ayant pas la nationalité française et vivant à l'étranger au dernier recensement) dans la population totale
Emploi	Actifs (personne ayant un emploi, chômeurs et militaires du contingent) dans la population totale Actifs dans la population totale masculine Actifs dans la population totale féminine Chômeurs dans la population active totale ⋆ Chômeurs étrangers dans la population active Chômeurs chez les actifs de 15-24 ans Chômeurs chez les actifs de plus de 50 ans Chômeurs dans la population active masculine Chômeurs dans la population active féminine Chômeurs depuis plus d'un an dans la population active totale Non-salariés (indépendants, employeurs, aides familiaux) dans la population active totale Emplois précaires (apprentis, intérimaires, stagiaires, CDD, emplois aidés) dans la population active totale Emplois stables (CDI, titulaires de la fonction publique) dans la population active totale Agriculteurs occupés dans la population active totale Cadres dans la population active totale ⋆ Ouvriers dans la population active totale ⋆

Toutes les variables de ce tableau sont exprimées en %

⋆ variable utilisée fréquemment dans la littérature

TABLE 5.2 – Description des variables socio-économiques construites à partir des données du recensement 1999, par domaine (suite)

Éducation et formation	Scolarisés chez les 6-15 ans
	En cours d'études parmi les plus de 15 ans
	Non diplômés (non en cours d'études) parmi les plus de 15 ans ⋆
	CAP-BEP (non en cours d'études) parmi les plus de 15 ans
	Bac-Brevet (non en cours d'études) parmi les plus de 15 ans ⋆
	Bac+2 (non en cours d'études) parmi les plus de 15 ans ⋆
	Diplômes supérieurs (non en cours d'études) parmi les plus de 15 ans
Logement et équipement	Maisons individuelles parmi les résidences principales
	Immeubles collectifs parmi les résidences principales
	Résidences principales dont les occupants ne sont pas propriétaires (locataires, sous-locataire ou logées gratuitement) ⋆
	Résidences principales de type HLM
	Résidences principales construites avant 1968
	Résidences principales construites après 1990
	Résidences principales de moins de 40m^2
	Résidences principales de plus de 150m^2
	Résidences principales sans baignoire ni douche
	Résidences principales sans WC
	Résidences principales sans chauffage
	Résidences principales avec un garage/box/parking
	Ménages sans voiture
	Ménages avec deux voitures ou plus ⋆

Toutes les variables de ce tableau sont exprimées en %

⋆ variable utilisée fréquemment dans la littérature

5.1.4 Les données socio-économiques de 2006 : Modifications par rapport à 1999

Bien que constituées en majorité de dénombrements identiques, les données issues du recensement 2006 présentent néanmoins certaines différences avec celles issues du recensement 1999 et par conséquent certaines variables construites pour ce dernier n'ont pas pu l'être pour 2006. *A contrario*, d'autres variables, similaires à celles qui existaient pour 1999 ou nouvelles, ont été construites. La table 5.3 résume ces différences.

On a pris le parti ici, comme précédemment, d'intégrer toutes les variables dont on estimait qu'elles avaient un lien avec le statut socio-économique et indifféremment de la sélection des variables de 1999. En effet, l'objectif de sélection des données demeure ici d'avoir un ensemble d'indicateurs aussi large que possible au sein duquel on sélectionnera et/ou synthétisera l'information.

TABLE 5.3 – Différences de variables utilisées entre les recensements 1999 et 2006

	Recensement 1999	Recensement 2006
Présentes en 1999 mais plus en 2006	Chômeurs étrangers	-
	Chômeurs depuis plus d'un an	-
	Résidences principales sans WC	-
	Migrants	-
Modifiées entre 1999 et 2006	Chômeurs de plus de 50 ans	Chômeurs de plus de 55 ans
	Résidences principales construites avant 1968	Résidences principales construites avant 1974
	Résidences principales de plus de 150m^2	Résidences principales de plus de 100m^2
Présentes en 2006 mais pas en 1999	-	Retraités chez les 15-64 ans
	-	Artisans occupés dans la population active totale
	-	Professions intermédiaires occupés dans la population active totale
	-	Employés occupés dans la population active totale
	-	Ménages ayant déménagé depuis moins de 2 ans
	-	Ménages ayant déménagé depuis moins de 5 ans
	-	Ménages ayant déménagé depuis plus de 10 ans
	-	Plus de 15 ans qui travaillent ailleurs que dans la commune de résidence dans la population des plus de 15 ans

Toutes les variables de ce tableau sont exprimées en %

5.2 Revenus fiscaux des ménages et enquête logement

Outre les résultats des recensements de 1999 et 2006, deux autres bases de données ont été utilisées. La première est la base de données des revenus fiscaux des ménages pour les années 2001 et 2006, et la seconde est l'enquête logement menée en 2001. Les indicateurs de revenu et de logement sont en effet très couramment utilisés dans la littérature et il semblait important de les inclure dans la sélection de variables de départ même si nous étions conscients que ces différentes bases couvraient des périodes différentes (ce point sera discuté dans la section 5.4).

Les revenus fiscaux des ménages donnent des indications à différentes échelles et unités sur la distribution des revenus au sein de la population. Ces bases sont fournies pour chaque année par l'INSEE et la Direction Générale des Impôts, via l'exploitation des déclarations de revenus et de la taxe d'habitation. Pour chaque zone géographique (avec des échelles allant de la France entière à l'IRIS) sont fournis différents quantiles de la distribution des revenus (quartiles, déciles) ainsi que la moyenne et l'écart-type du revenu. L'indice de Gini, qui vise à quantifier les inégalités de revenus, est également fourni.

Par ailleurs, ces indicateurs sont disponibles à différents niveaux d'observation : personne, ménage ou unité de consommation. C'est ce dernier niveau d'observation qui a été retenu et qui est préconisé par l'INSEE [106] lorsque l'on souhaite effectuer des comparaisons de revenu entre différentes zones, même lorsque la composition de leurs ménages est différente. Le revenu fiscal par unité de consommation est « le revenu du ménage rapporté au nombre d'unités de consommation qui le composent ». L'unité de consommation du ménage est définie de la manière suivante : le premier adulte du ménage compte pour 1 unité de consommation ; les autres personnes de 14 ans ou plus comptent pour 0,5 ; et les enfants de moins de 14 ans comptent pour 0,3.

Dans la mesure où il n'y avait pas de données disponibles sur le revenu pour 1999, ce sont celles de l'année 2001 qui ont été utilisées avec les données du recensement 1999. Les données du recensement 2006 ont quant à elles été reliées aux données de revenu 2006. C'est dans les deux cas le revenu médian par unité de consommation qui a été conservé car il s'agit d'un indicateur très fréquemment utilisé dans de nombreux domaines et bien connu de tous. Des critères de confidentialité des données (nombre de ménages ou d'habitants minimaux dans une zone nécessaire à la publication des résultats pour celle-ci) causent néanmoins l'apparition d'un certain nombre de valeurs manquantes pour ces indicateurs. La manière dont celles-ci ont été traitées est détaillée dans la section suivante.

L'enquête logement, quant à elle, est l'une des plus anciennes enquêtes menées par l'INSEE (elle est réalisée depuis presque 60 ans) et l'une des plus vastes en taille d'échantillon (plusieurs dizaines de milliers de logements). Elle vise à décrire le parc des logements, les conditions d'occupation des résidences principales par les ménages ou encore le coût du logement pour ceux-ci. Les données de cette enquête pour l'année 2001 ont été utilisées pour ajouter aux variables du recensement 1999 une variable indiquant le nombre moyen de personnes par pièce des résidences principales et la proportion de résidences principales comptant plus d'une personne par pièce dans les IRIS. Ces variables font en effet partie des indicateurs qui sont fréquemment utilisés dans la littérature, y compris dans les indices socio-économiques déjà existants (voir section 7.1). Faute d'une base de données similaire disponible pour 2006, ces deux variables n'ont pas été inclues pour cette année.

5.3 Traitement des données manquantes

Malgré la qualité des données, certains IRIS présentent des données manquantes. Une faible proportion d'IRIS n'a absolument aucune donnée pour les variables issues du recensement. L'une des raisons de cette absence totale de donnée est la confidentialité qui est imposée à l'INSEE de ne pas divulguer les données d'un IRIS si celui-ci n'a pas un nombre minimum d'habitants ou de ménages, garantissant ainsi l'anonymat des personnes. Dans ce cas, les valeurs de ces IRIS n'ont pas été traitées ni complétées et ont été laissées manquantes, ce qui excluait d'office ces IRIS de l'analyse.

Cependant, la majorité des valeurs manquantes proviennent du revenu médian. Plusieurs raisons sont avancées pouvant expliquer la présence de données manquantes relatives au revenu médian. Comme précédemment explicité, afin de respecter les accords CNIL, le revenu ne peut être communiqué sur un IRIS trop peu peuplé.

Une seconde explication est liée à la nature des bases disponibles auprès de l'INSEE. Certaines communes n'étant composées que d'un seul IRIS, l'information sur le revenu est

présente dans les bases à l'échelle de la commune mais pas dans celles à l'échelle de l'IRIS. Dans ce cas, la valeur du revenu médian renseignée dans la base à l'échelle communale a été répercutée sur celle à l'échelle de l'IRIS.

La valeur du revenu médian restait manquante pour certains IRIS (15 à 25% suivant les agglomérations), on a par conséquent souhaité compléter celle-ci afin de ne pas devoir exclure ces IRIS également et réduire fortement la zone d'étude (et le nombre d'événements sanitaires déjà faible). En effet, l'absence d'une seule valeur pour un IRIS implique automatiquement que les indicateurs composites construits pour celui-ci seront également manquants (cet IRIS ne contribuera donc pas à l'analyse), alors même que les autres variables servant à calculer ces indicateurs peuvent être connues pour l'IRIS.

De plus, la structure spatiale des données et les techniques mises en œuvre pour la prendre en compte peuvent également être impactées par la présence de « trous » dans la zone d'étude (voir section 10.2), ce qui se produirait dans le cas de valeurs manquantes.

Différentes techniques ont donc été étudiées pour compléter les valeurs manquantes du revenu médian. Parmi elles, nous avons envisagé de chercher à prédire le revenu médian grâce aux autres variables socio-économiques renseignées à l'aide d'un modèle de régression linéaire. Cependant, une telle prédiction impliquerait d'introduire de la colinéarité entre le revenu médian et les autres variables, ce qui nous a paru préjudiciable. En effet, l'objectif ici est d'éliminer les corrélations et colinéarités entre nos variables explicatives (voir partie II).

La méthode suivante a été appliquée pour compléter les valeurs de revenu médian manquantes :

- Si la commune est composée de **3 IRIS ou moins**, tous sans valeur, chaque IRIS de la commune prend la valeur du revenu médian de la commune. On a en effet considéré que si la commune ne comptait que 2 ou 3 IRIS tous sans valeur alors celle-ci était de taille suffisamment petite pour estimer la valeur manquante par le revenu médian de la commune elle-même.
- Si la commune est constituée de **plus de 4 IRIS** sans valeur ou si plus de 3 IRIS dans une commune sont sans valeur, chaque IRIS sans valeur prend la valeur moyenne des revenus médians des IRIS voisins (adjacents). On a ici considéré que si une commune avait plus de 4 IRIS, alors elle était de taille suffisamment importante pour présenter une hétérogénéité de revenus médians en son sein qui justifiait de prendre en compte les valeurs des unités voisines.

Afin d'évaluer l'ajout d'auto-corrélation spatiale provoqué par cette technique, nous avons utilisé un indicateur classique : l'indice de Moran [107].

Indice de Moran

Soit :

- $z_i = x_i - \bar{x}$ la déviation à la moyenne d'une variable x pour une unité i,
- $w_{i,j}$ le poids spatial (basé sur la distance) entre deux unités i et j,
- n le nombre d'unités,
- $P = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}$,

alors l'indice de Moran est défini comme :

$$I = \frac{n}{P} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

L'interprétation de cet indice qui varie entre -1 et 1 permet d'avoir une indication sur le degré d'auto-corrélation spatiale de la variable x dans la zone. Ainsi, plus l'indice est proche de 1 et plus il y a de similarités entre unités voisines et respectivement, plus il est proche de -1 et plus il y aura de dissimilarités entre unités voisines. Un indice de 0 indique une distribution spatiale aléatoire.

On constate ici qu'une auto-corrélation spatiale est déjà présente initialement dans les données (de revenu notamment), ce qui nous assure que notre traitement des données manquantes ne va pas en introduire là où il n'en existait pas et n'en ajoutera qu'une faible proportion (comme on peut le voir table 5.4).

TABLE 5.4 – Indice de Moran du revenu non complété et complété dans les différentes agglomérations pour 2006

	Revenu non complété	Revenu complété
Paris - Petite Couronne	0,60	0,63
Aix-Marseille	0,39	0,44
Grand Lyon	0,29	0,32
Lille Métropole	0,31	0,36

Ces valeurs sont toutes significativement différentes de 0

Ainsi, on obtient pour chaque IRIS une valeur de revenu médian, ne dépendant pas directement des autres variables socio-économiques tout en demeurant plausible vis-à-vis des autres données disponibles.

Notons qu'en plus de l'indice de Moran, nous avons également testé et comparé la procédure de création d'indices socio-économiques exposée en section 7.2 à la fois sur les données où les valeurs de revenu médian manquantes sont complétées et celles non complétées. Nous avons alors pu constater que les résultats obtenus ne présentaient que de faibles différences en termes de sélections de variables, de contributions et de corrélations de celles-ci avec l'indice obtenu. Les tests statistiques (test de comparaison de moyenne et/ou test de Wilcoxon pour échantillons appariés) indiquent cependant une différence significative entre les valeurs de l'indice créé avec les données complétées et celles de l'indice créé avec les données non complétées, qui semble pouvoir s'expliquer par l'ajout des IRIS dont les valeurs de revenu étaient manquantes (et donc l'ajout de toutes leurs caractéristiques socio-économiques) dans la procédure.

L'ensemble de ces différents éléments nous a conduit à utiliser systématiquement les données avec revenu complété par la suite afin d'avoir un maximum d'IRIS ayant une valeur de l'indice composite construit (et donc d'éviter les « trous » dans la zone d'étude lors des analyses suivantes).

5.4 Discussion

Pour conclure, on utilisera donc deux ensembles de variables, l'un pour l'année 1999 et l'autre pour 2006. L'ensemble des variables considérées pour l'année 1999 est celui présenté dans les tables 5.1 et 5.2 auquel on ajoute le nombre moyen de personnes par pièce et la proportion de logements avec plus d'une personne par pièce issus de l'enquête logement de 2001, et le revenu médian 2001.

Pour l'année 2006, on considérera les mêmes variables issues du recensement mais avec les modifications présentées dans la table 5.3, auquelles on ajoute le revenu médian 2006.

Nous avons volontairement gardé à cette étape un nombre important de variables afin de laisser les techniques utilisées par la suite sélectionner et/ou synthétiser celles-ci. On s'est donc contenté ici de critères basés uniquement sur la littérature relative à la notion de statut socio-économique. Ces variables sont construites directement à partir des données du recensement, ce qui devrait garantir leur bonne qualité.

Dans les quelques cas où il y avait des données manquantes, celles-ci ont été complétées en tirant parti de la structure spatiale des données. Bien que ce procédé ajoute un peu d'autocorrélation spatiale (table 5.4), nous avons préféré ceci à l'ajout d'une colinéarité entre les variables.

Une limite aux données choisies demeure les différences d'année de référence entre certaines bases. Ainsi, nous avons parfois utilisé simultanément des données de 1999 et d'autres de 2001. Ceci pourrait poser quelques problèmes dans le cas d'IRIS ayant connus d'importantes évolutions entre ces deux années. En pratique et sur les zones concernées, l'étude des corrélations et des nuages de points entre variables de différentes années (par exemple la comparaison entre le revenu médian et la proportion de cadres, de chômeurs ou de bac+2) ne nous permet pas de distinguer d'incohérence ou de contradiction flagrante entre ces données.

Chapitre 6

Données d'expositions environnementales

Dans ce chapitre, nous présenterons les différentes nuisances environnementales qui ont été incluses dans l'étude et la manière dont celles-ci ont été prises en compte. Cette prise en compte à l'échelle de l'IRIS fait par ailleurs partie des difficultés qui ont pu survenir durant la construction des variables : en effet, les échelles habituelles de mesure ou de modélisation des institutions et organismes compétents ne coïncidaient pas forcément avec l'échelle de l'IRIS. Le but a donc été d'obtenir des indicateurs donnant des informations aussi cohérentes, fiables et pertinentes que possible sur les niveaux de nuisances ou sur la population exposée dans un IRIS donné.

6.1 La pollution atmosphérique : l'indicateur dioxyde d'azote

6.1.1 Justification du choix du NO_2

Au sein de la très vaste et très hétérogène famille des polluants de l'air, le dioxyde d'azote (NO_2) a été choisi comme principal polluant de l'air d'intérêt. Le dioxyde d'azote est un polluant secondaire issu du monoxyde d'azote (NO) qui est lui-même produit principalement par les processus de combustion à haute température des énergies fossiles, et en particulier par le trafic routier (notamment les moteurs diesel) dans les contextes urbains. Le NO_2 est par ailleurs un précurseur d'autres types de polluants de l'air, différentes actions chimiques pouvant le transformer en ozone (O_3) ou le faire entrer dans la catégorie des particules fines (PM). Il peut se transformer par ailleurs en acide nitrique (HNO_3) aussi bien dans l'atmosphère, causant ainsi des pluies acides, qu'en entrant en contact avec les muqueuses pulmonaires lorsqu'il est inhalé, ce qui cause des irritations des bronches [108].

De manière générale, le NO_2 est considéré comme un bon indicateur de la pollution de l'air urbain générée par le trafic. Par ailleurs, ce polluant présente une hétérogénéité spatiale plus importante que d'autres, ce qui permet d'envisager des contrastes selon les caractéristiques socio-économiques du territoire. Enfin, l'association entre l'exposition aussi bien à court qu'à long terme au NO_2 et la santé est de plus en plus clairement établie, au moins si l'on considère ce polluant comme un indicateur caractérisant le mélange complexe issu des mêmes sources [109].

6.1.2 Sources des données

Les Associations Agréées de Surveillance de la Qualité de l'Air (AASQAs) sont des organismes chargés par l'État de mesurer, étudier et communiquer sur la qualité de l'air ambiant en France. Chaque région compte ainsi au moins une AASQA et, pour les agglomérations étudiées au sein du projet, ce sont les associations AirParif [110], Atmo PACA [111], Air Rhône-Alpes [112] et Atmo Nord-Pas-de-Calais [113] qui effectuent cette mission. Leurs données et mesures sur un long historique ainsi que leur savoir-faire et leurs compétences en matière de pollution de l'air et modélisation de celle-ci ont fait d'elles des partenaires de choix pour prendre en compte cette nuisance environnementale.

Ces mesures peuvent être réalisées par plusieurs types de stations, automatiques, semi-automatiques ou même mobiles, suivant le type de polluant étudié. On distingue principalement deux types de stations en situation urbaine : les stations de fond, éloignées des voies de circulation, et les stations de trafic [114].

Dans le cadre du partenariat entre l'EHESP et les AASQA autour du projet Equit'Area, ces dernières ont mis à disposition l'ensemble des mesures disponibles issues de leurs réseaux de mesures, de 2002 à 2009 et sur les zones d'études concernées. Leur expertise a également servi à modéliser ces mesures de manière à obtenir, au final, les concentrations annuelles de NO_2 par IRIS.

6.1.3 Modélisation du NO_2

L'évaluation des niveaux de pollution dans les IRIS nécessite plusieurs phases de modélisation à partir des mesures du NO_2 d'une part, mais aussi d'un grand nombre d'autres données comme des informations météorologiques, des niveaux de pollution « de fond », ou encore des émissions de trafic, surfaciques et ponctuelles.

Les mesures directes n'étant pas forcément réalisées partout pour toutes les années, une première phase de modélisation consiste dans la reconstitution des concentrations annuelles manquantes avec des adaptations géostatistiques et des interpolations spatiales. Une autre partie importante de modélisation réalisée par les AASQA consiste à établir des cartes de concentration à l'échelle de l'IRIS à partir des différentes données évoquées ci-dessus.

De nombreux modèles numériques différents sont utilisés [115–118] mais leurs données d'entrée sont systématiquement de même type (météorologie, pollution de fond, cadastre d'émissions de polluants, ...) et ils permettent tous d'obtenir une cartographie fine des concentrations, ici du NO_2, avec des résolutions spatiales allant de 10 à 50m. Il est à noter que ces modèles numériques sont « calés » sur les mesures effectuées par les stations.

Le principal travail de modélisation effectué par les AASQA dans le cadre du projet a été l'élaboration d'une méthode d'agrégation des résultats par IRIS pour produire différents indicateurs : NO_2 annuel moyen par IRIS, écart-type des valeurs de NO_2 par IRIS, estimation de la population exposée à la valeur limite (>40 $\mu g/m^3$) par bâtiment et par IRIS. De plus, une classification ascendante hiérarchique (CAH) des concentrations moyennes de NO_2 a permis de déterminer pour chaque IRIS la station représentant le mieux l'évolution journalière de ses concentrations.

> Dans le cadre de cette thèse, *seules les concentrations moyennes annuelles ont été considérées*, principalement parce que toutes les données n'étaient pas disponibles pour l'ensemble des agglomérations. De plus, ce choix de ne retenir que l'estimation ponctuelle (ici la concentration moyenne) se justifie pleinement dans le contexte de santé publique dans lequel se positionne ce travail. En effet, la santé publique s'intéresse à l'exposition de populations et cherche par conséquent à avoir un « profil moyen » de l'exposition des IRIS concernant un maximum de personnes. Enfin, la concentration annuelle moyenne est l'un des indicateurs les plus couramment utilisés dans ce domaine, pour lequel des valeurs de référence sont disponibles, des recommandations sont formulées et une importante littérature existe.

L'utilisation des autres indicateurs, en particulier l'écart-type des concentrations par IRIS, pourrait néanmoins permettre de « nuancer » les informations apportées par la moyenne en donnant, par exemple, plus de poids aux IRIS dont les concentrations sont plus précises ou en donnant une mesure de la « fiabilité » des informations apportées par la moyenne dans l'interprétation des résultats des analyses. L'utilisation d'une pondération de la moyenne par l'inverse initialement envisagée a dû être écartée et n'a pas été explorée car toutes les expositions environnementales ne permettaient pas d'estimer des mesures de dispersion.

6.2 Les nuisances sonores

6.2.1 Justification du choix des nuisances sonores

Jusqu'à une période récente, peu d'études ont exploré le rôle de l'exposition au bruit dans les inégalités sociales de santé, ou comme un facteur de risque sur la santé de manière générale. Cependant, il est désormais de plus en plus établi que l'exposition aux nuisances sonores est très inégale entre les différents groupes socio-économiques.

Selon l'Organisation Mondiale de la Santé, les effets sur la santé d'une exposition à long terme à un bruit trop important incluent des troubles du sommeil, des troubles cardiovasculaires, des troubles mentaux, des problèmes de productivité au travail ou de performances d'apprentissage, ou encore des problèmes auditifs ou des acouphènes. Des effets sur le stress ou des issues de grossesse défavorables ont également été observés [119].

Pour ces différentes raisons, l'exposition au bruit (ici principalement issu du trafic routier) a donc été incluse dans cette étude pour le projet Equit'Area.

6.2.2 Sources des données et modélisation des nuisances sonores

L'établissement de cartographies du bruit dans les agglomérations de plus de 100 000 habitants résulte de la directive européenne 2002/49/CE du 22 juin 2002 [120]. Celle-ci a pour but de permettre une évaluation harmonisée de l'exposition au bruit dans l'environnement (principalement le bruit des trafics aérien, routier et ferroviaire, ainsi que le bruit industriel) dans la Communauté Européenne, de mettre en œuvre des actions visant à réduire cette exposition et d'informer le public. Elle donne par ailleurs une méthodologie commune pour la modélisation des expositions afin d'estimer la valeur maximale des

niveaux de bruit calculés sur chaque face des bâtiment à 2m en avant de la façade et à 4m du sol.

L'estimation des niveaux de bruit est réalisée en deux étapes :

1. **La modélisation des niveaux de bruit à l'échelle des bâtiments.** Ceux-ci sont obtenus par modélisation à partir des caractéristiques des sources émettrices, de la topographie, des conditions météorologiques ... La modélisation acoustique (faisant l'objet de la directive européenne citée précédemment) donne alors une estimation des niveaux de bruit à 4 mètres de haut avec une résolution spatiale de 10×10 mètres, qui sert alors à attribuer à chaque bâtiment une valeur de bruit.

2. **L'estimation des niveaux d'exposition des IRIS.** La méthodologie de construction d'indicateurs d'exposition au bruit à l'échelle de l'IRIS a été mise au point sur l'agglomération de Lyon en collaboration avec le Centre Scientifique et Technique du Bâtiment (CSTB ; cf infra)[121] . Cette estimation est basée sur les niveaux de bruit de différentes sources et périodes (6h-18h, 18h-22h, 22h-6h, ...) par bâtiment, la nature (habitation, industrie, commerce, ...) et le volume de ces derniers.

Dans la mesure où chaque IRIS compte plusieurs centaines, voire milliers de bâtiments (en moyenne, 300 valeurs de niveau de bruit-bâtiment par IRIS), l'agrégation de ces données est nécessaire pour obtenir des indicateurs à l'échelle de l'IRIS. Le CSTB de Lyon, partenaire du projet, a pour cela choisi de ne retenir qu'un type de valeur de niveau de bruit : l'indicateur L_{DEN} total. Cet indicateur représente une mesure de l'exposition au bruit de toutes sources sur 24h. Il prend en compte les bruits émis tout au long de la journée mais aussi en soirée et la nuit en donnant une pénalité aux bruits émis le soir et la nuit pour tenir compte de la différence de sensibilité des personnes à ces moments. Il est exprimé en décibels acoustiques (dB(A)).

À partir de cet indicateur de base disponible pour chaque bâtiment, trois types d'indicateurs de niveaux de bruit à l'échelle vont être construits. Dans chaque cas, des indicateurs de tendance centrale (moyenne arithmétique, énergétique, médiane) et de dispersion (quantiles 5, 10, 20, 25, 75, 80, 90 et 95) des distributions de niveaux de bruit au sein de chaque IRIS sont construits. Ces trois types d'indicateurs sont les suivants :

- Les indicateurs de « phase 1 » ne tiennent compte que des valeurs de bruit par bâtiment au sein des IRIS, sans tenir compte des types de bâtiment ou de leur population.

- Les indicateurs de « phase 2 » se basent sur la première phase mais prennent en outre en compte le nombre d'habitants de chaque bâtiment, mais pas son type. La population de chaque bâtiment d'habitation est évaluée au prorata du volume de ce bâtiment par rapport à la population totale dans l'IRIS. Une pondération est alors introduite dans les calculs des moyennes.

- Les indicateurs de « phase 3 » se basent sur ceux de la deuxième phase, mais prennent en compte le type de bâtiment. Notamment, ils ajoutent à la pondération par la population des bâtiments d'habitation la prise en compte des bâtiments de type commerces et bureaux pour mieux évaluer le bruit subi par les personnes dans ces bâtiments durant la journée.

> Nous disposons donc de 11 indicateurs pour chacune des 3 phases fournis par le CSTB. Dans ce travail, *nous avons choisi de nous concentrer sur les indicateurs de tendance centrale uniquement*, pour les raisons précédemment exposées dans le cas du dioxyde d'azote en section 6.1.3. Un développement futur pourrait donc être réalisé afin d'inclure également les différents indicateurs de dispersion disponibles, pour pondérer les indicateurs de tendance centrale ou encore faire des comparaisons directement sur les distributions de bruit des IRIS.

Notons par ailleurs qu'après avoir effectué des analyses préliminaires sur ces 33 variables, il est apparu que les corrélations entre les indicateurs de phase 2 et ceux de phase 3 étaient extrêmement importantes (toutes les corrélations deux à deux entre indicateurs similaires sont supérieures à 0,92). Ceci nous a conduit à écarter les indicateurs de phase 3 de la suite des analyses puisqu'ils sont plus complexes à construire et nécessitent plus de données sans pour autant apporter d'information supplémentaire par rapport à ceux de phase 2.

6.3 Les indicateurs de proximité : industries, axes routiers et espaces verts

Outre les indicateurs quantitatifs propres à des niveaux agrégés de pollution de l'air ou de bruit à l'échelle de l'IRIS, des indicateurs dits « de proximité » ont également été inclus. Certaines nuisances ont en effet des origines ponctuelles ou de surfaces précises, dont les effets sont particulièrement dépendants de la distance à laquelle elles se trouvent des populations. Pour prendre en compte ces nuisances, différents types d'indicateurs visant à refléter la proximité entre les IRIS et ces sources ont été inclus.

6.3.1 Les industries polluantes

La prise en compte des industries polluantes apparaît comme nécessaire et importante lorsqu'il s'agit d'étudier l'impact de l'environnement sur les inégalités sociales de santé. Outre l'impact direct sur la santé qu'elles peuvent avoir en tant que sources de polluants aussi bien dans l'air, l'eau que les sols [122], elles peuvent également avoir un impact sur l'état psychologique des populations et la santé perçue [123]. Par ailleurs, comme on l'a détaillé dans la section 1.2.2, la proximité aux industries polluantes peut fortement varier suivant les groupes socio-économiques, en raison de la forte hétérogénéité de la distribution spatiale des installations polluantes [66, 70].

L'évaluation de la proximité aux industries polluantes des IRIS a donc un intérêt particulier lorsque l'on souhaite étudier les liens entre la santé, le statut socio-économique et les expositions environnementales. Afin de prendre en compte ces industries polluantes, c'est la base de données issue du Registre Européen des Rejets et Transferts de Polluants (E-PRTR, pour European Pollutant Release and Transfer Register) qui a été utilisée [124]. Ce registre, qui découle directement depuis 2007 de l'ancien Registre Européen des Émissions Polluantes (EPER, établi en 2000) est le plus important en Europe fournissant des données sur les émissions industrielles de l'ensemble de l'Union Européenne. Il contient des informations sur environ 28 000 industries recouvrant une soixantaine d'activités économiques différentes et, pour chaque établissement, fournit des informations annuelles sur les quantités de rejets de polluants (pour un total de 91 polluants concernés) dans l'air, l'eau et le sol ainsi que le transfert de déchets hors du site. Ces quantités sont des quantités déclarées

par les industries elles-mêmes lorsque leurs émissions dépassent le seuil réglementaire d'au moins l'un des 91 polluants. Il s'agit donc des installations les plus polluantes.

Dans ce travail, seules les émissions dans l'air ont été prises en compte pour les industries dans un premier temps. En effet, ce sont ces émissions qui sont les plus susceptibles d'avoir un effet sur la mortalité infantile. Les industries qui ont été retenues sont ainsi celles qui ont dépassé au moins une année (durant la période 2004-2009, les données pour les années précédentes n'étant pas disponibles) le seuil d'au moins un polluant de l'air renseigné dans l'E-PRTR. Cependant, les données géographiques fournies dans ce registre n'étant parfois pas de bonne qualité, une vérification a été réalisée pour l'ensemble des sites retenus avec, si nécessaire, la recherche des coordonnées géographiques exactes afin de géolocaliser le plus correctement possible les sites industriels et les sources réelles de pollution. Une fois cette vérification et les corrections faites, une base de données géographique des industries polluantes a donc été établie, à partir de laquelle des indicateurs de proximité ont été construits.

Comme il s'agit de sources « ponctuelles » possédant une zone d'effet plus ou moins compliquée à déterminer, de nombreux types d'indicateurs sont possibles pour refléter la proximité aux industries à l'échelle de l'IRIS. Ici, **deux types d'indicateurs ont été utilisés** :

- **Les indicateurs basés sur la présence/absence ou le nombre d'industries** dans l'IRIS. Ce sont les plus simples à construire.
- **Les indicateurs basés sur l'estimation d'une zone d'influence autour de l'industrie**, souvent appelée *buffer*, représentant par exemple l'aire de dispersion des polluants autour d'une cheminée. La distinction est alors faite entre les IRIS sous l'influence de l'industrie et ceux n'étant pas sous son influence. Un grand nombre de sous-types d'indicateurs existent suivant la manière dont est déterminée si un IRIS est sous l'influence d'une industrie ou non.
  Dans le cadre de ce travail, les indicateurs de ce type sont la présence/absence de buffers, et le nombre de buffers, en considérant la présence comme la simple intersection entre le buffer et la frontière de l'IRIS. La direction des vents n'a pas été prise en considération pour caractériser les buffers qui sont à profil circulaire. Le rayon des buffers choisi a été de 500m ou de 1km.

D'autres indicateurs ont été envisagés (buffers de rayon proportionnel aux émissions, par exemple), mais les émissions présentes dans la base de données utilisée étant déclarées par les industries elles-mêmes (et parfois données dans des unités différentes), la fiabilité de cette information nous a semblé trop faible. Nous avons également étudié la possibilité de construire des indicateurs spécifiques à chaque type de polluant de l'air (voire à chaque polluant de l'air) mais le nombre d'industries dans les zones d'études s'est révélé insuffisant pour cela. C'est aussi pour cette raison que nous avons considéré ces indicateurs (y compris ceux de nombre d'industries ou de buffers) comme des variables qualitatives, dont nous avons recodé les modalités lorsque cela s'avérait nécessaire afin d'avoir un nombre suffisant d'IRIS dans chaque.

> Ce choix d'utiliser des variables qualitatives nous imposera cependant d'utiliser par la suite des techniques adaptées pour traiter à la fois des variables quantitatives

et qualitatives afin de répondre à l'objectif de construction d'indicateurs composites (voir chapitre 8).

6.3.2 Les axes routiers à forte densité de trafic

Le trafic automobile est une source importante de nuisances environnementales, qui implique aussi bien la pollution de l'air que le bruit, mais elle est aussi relativement localisée. En effet, ce sont surtout les voies ayant un trafic important qui contribuent à ces nuisances, et la littérature montre que les concentrations de polluants de l'air liés au trafic diminuent très fortement avec la distance (elles sont ainsi divisées par quatre à 100m d'un axe routier, par huit à 200m et par plus de dix à 300m pour des polluants comme le CO ou le NO_2)[125]. Par ailleurs, différents tronçons d'un même axe peuvent ne pas avoir les mêmes trafics et par conséquent représenter une exposition différente.

Un premier type de variables considéré a été les indicateurs de présence/absence d'une voie avec un trafic dépassant un certain seuil. Dans notre cas, deux seuils ont été testés, plus de 5 000 véhicules par jour et plus de 10 000 véhicules par jour suivant les conseils d'experts locaux. Cependant, ces indicateurs se sont révélés trop « grossiers » dans notre cas puisque la très grande majorité des IRIS des zones étudiées possèdent au moins une voie de trafic au-dessus de ces seuils. L'intérêt de ces variables étant très limité à cause de leur très faible variabilité dans les zones d'études, elles n'ont pas été conservées.

Ceci implique qu'il est nécessaire d'utiliser des indicateurs permettant de prendre en compte finement la population à proximité des voies à fort trafic et qu'une simple indication de présence ou d'absence de voie à fort trafic dans un IRIS est peu représentative de l'exposition réelle de la population.

Nous avons donc cherché à construire un autre type d'indicateur de trafic, à partir des données de modélisations du trafic routier effectuées par les instances locales. Celles-ci donnent, par tronçon routier, les estimations de trafic en nombre de véhicules par jour qui sont notamment utilisées dans les modèles d'estimation de la pollution de l'air comme source d'émission. Les données de bâti, fournies par l'Institut national de l'information géographique et forestière (IGN)[126], et celles de population issues du recensement sont également utilisées.

À partir de ces données, une famille d'indicateurs de proportion de la population exposée au trafic a été construite. Les tronçons à « fort » trafic considérés sont ceux ayant plus de 5 000 véhicules/jour. Autour de ces tronçons, une bande « tampon » a été construite, avec comme distances autour de l'axe 100m, 150m, 200m, 250m ou 300m. Une fois ces bandes constituées, les bâtiments habitables renseignés dans les données de bâti et localisés dans les bandes ont été identifiés. Enfin, la population présente dans ces bâtiments a été estimée sur la base de leur volume et des données de population issues des recensements 1999 et 2006.

On obtient ainsi pour chaque largeur de bande et les deux années de recensement une estimation de la proportion de la population de chaque IRIS qui est exposée à un trafic important. Ceci permet une évaluation plus fine de l'exposition au sein des IRIS en prenant en compte l'aspect très localisé de l'exposition au trafic routier.

6.3.3 Les espaces verts

Les espaces verts (que l'on peut définir comme les zones ouvertes non développées avec de la végétation naturelle, les parcs ou les forêts) et la proximité à ceux-ci, contrairement aux autres expositions présentées ici, constituent une source « d'exposition positive ». En effet, une littérature récente tend à montrer que les espaces verts ont des effets bénéfiques aussi bien sur de nombreux événements sanitaires (maladies cardiovasculaires, pression artérielle, surpoids et obésité, stress, santé mentale et émotionnelle, bien-être) que sur la mortalité et l'espérance de vie [127, 128]. Les mécanismes derrière ces effets ne sont pas encore bien établis, mais l'augmentation de l'activité physique et de la marche, les nuisances environnementales et sonores réduites ou encore l'augmentation des liens sociaux ont été évoquées. Par ailleurs, l'accès et la proximité des espaces verts diffèrent suivant les groupes socio-économiques, ce qui peut donc conduire à un effet sur les inégalités sociales de santé [127]. Tout ceci justifiait donc l'inclusion de cette exposition positive dans le projet Equit'Area.

Les données utilisées pour prendre en compte cette exposition positive proviennent de l'imagerie satellite, et plus précisément de la base de données Corine Land Cover [129]. Cette base de données européenne est pilotée par l'Agence européenne de l'environnement, et sa partie française est réalisée par le Service de l'Observation et des Statistiques du Ministère de l'Écologie, du Développement Durable et de l'Énergie. À partir d'images satellites, une interprétation humaine est effectuée pour déterminer l'occupation biophysique des sols suivant une nomenclature précise (en trois niveaux, comptant 5 catégories au niveau le plus élevé et 44 au niveau le plus fin). Les bases de données établies sont ensuite librement disponibles.

À partir de ces données, deux indicateurs d'espaces verts ont été construits : la surface verte en valeur absolue et la proportion de surface verte au sein d'un IRIS. Ces indicateurs permettent ainsi d'avoir à la fois des informations sur l'effet absorbant des espaces verts (qui dépend surtout de la taille réelle de ceux-ci) mais aussi sur l'environnement interne des IRIS et la proximité des populations aux espaces verts dans leur lieu de vie.

6.4 Discussion

Dans ce chapitre, nous avons présenté les différentes variables environnementales qui seront considérées par la suite (en particulier dans le chapitre 8).

Le choix que nous avons effectué ici a été de ne considérer que des indicateurs de tendance centrale dans le cas des niveaux de bruit et de concentration de dioxyde d'azote, alors que d'autres indicateurs (de dispersion) étaient également disponibles. En effet, ce travail s'inscrivant dans un projet de santé publique, il est important que les variables environnementales retenues reflètent le mieux possible l'exposition des populations résidant dans les IRIS, plutôt que des expositions touchant une faible proportion de personnes.

Nous avons également choisi, suite aux analyses descriptives de considérer les indicateurs de proximité aux industries polluantes comme des variables qualitatives, bien que ce choix nous conduise à utiliser par la suite des techniques spécifiques.

Parmi les limites des données considérées ici, la plus importante est l'hétérogénéité des sources de données et, par conséquent, des fiabilités des indicateurs. Nous mélangeons ainsi des variables issues de modélisations, de bases de données réglementaires ou encore d'imagerie satellite, qui peuvent donc comporter (ou non) une incertitude sur la modélisation, les données d'entrée, les déclarations faites, ... Ceci pourrait éventuellement biaiser les résultats dans le cas de modèles explicatifs intégrant directement toutes ces variables. Néanmoins, ces incertitudes devraient avoir un impact moindre dans ce présent travail puisque l'objectif ici est de synthétiser les informations de l'ensemble de ces variables en un indicateur qualitatif (voir chapitre 8).

Conclusion de la première partie

Dans cette partie, nous avons présenté les différentes zones d'études, l'unité statistique et géographique ainsi que les différentes variables sanitaires, socio-économiques et environnementales choisies qui seront utilisées par la suite.

Ce sont ainsi les quatre plus grandes agglomérations françaises de Paris et sa Petite Couronne, Marseille, Lyon et Lille qui seront étudiées afin d'avoir des contrastes vis-à-vis des caractéristiques socio-économiques et environnementales. L'étude sera réalisée à l'échelle de l'IRIS, niveau qui correspond globalement au quartier, et qui permet de minimiser le biais écologique et d'avoir une interprétation fine des résultats.

Sur chacun des IRIS, nous disposons de données de mortalité infantile et néonatale issues des registres de décès des communes, et des données de naissances obtenues ou estimées à partir du recensement de la population.

Les recensements de 1999 et 2006 sont aussi la source des données socio-économiques, qui sont représentées par une cinquantaine de variables donnant des informations sur les domaines de la démographie, de l'emploi, du logement, de la mobilité ou de l'éducation pour ces deux années.

Les variables environnementales enfin, sont issues en majorité de l'expertise des partenaires du projet et de bases de sources diverses (mesures directes, modélisations, registres européens, imagerie satellite) et nous permettent d'avoir des informations sur la pollution de l'air, les niveaux de bruit, la proximité aux industries polluantes, la proportion de population exposée au trafic ou encore la proximité aux espaces verts.

L'ensemble de ces données sera utilisé dans la suite pour mener à bien autant les objectifs de la thèse que ceux du projet de recherche dans lequel elle s'inscrit.

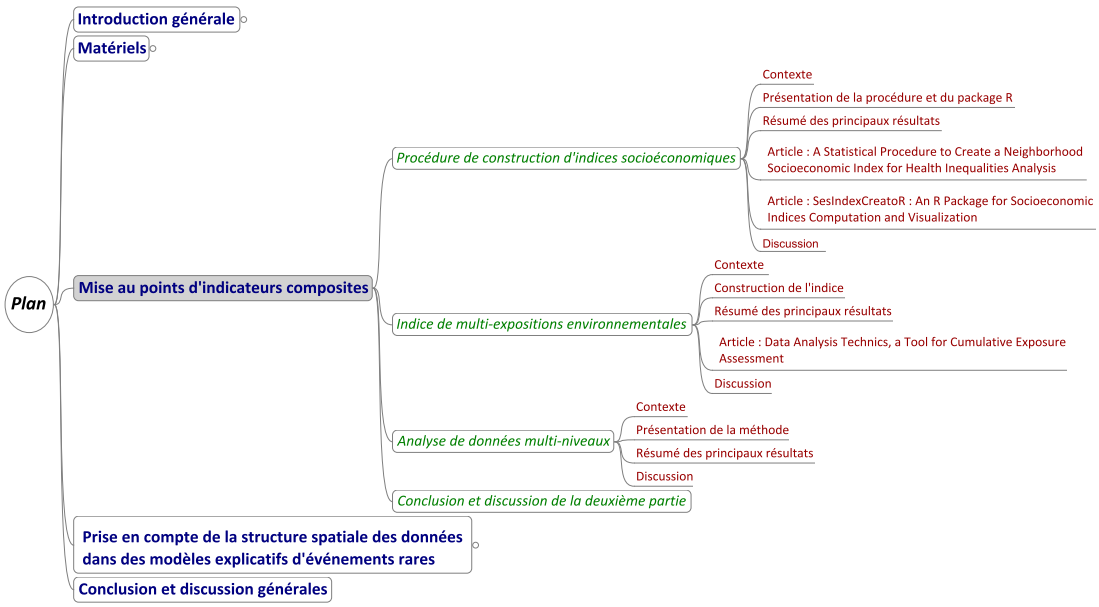
Deuxième partie

Mise au point d'indicateurs composites

Mise au point d'indicateurs composites

Dans cette seconde partie nous détaillerons le premier des objectifs principaux de cette thèse, qui consiste en la mise au point d'indicateurs composites. Tout d'abord, nous exposerons dans le chapitre 7 le développement d'une procédure de construction d'indices socio-économiques et la création d'un package du logiciel R l'implémentant, puis nous détaillerons au chapitre 8 la création d'un indice de multi-expositions environnementales, et nous terminerons au chapitre 9 en évoquant une technique d'analyse de données multi-niveaux que nous avons exploré comme perspective.

Pour chacun de ces chapitres, il sera d'abord expliqué le contexte dans lequel le travail s'insère. Les aspects méthodologiques et mathématiques seront détaillés ensuite avant de résumer les principaux résultats obtenus et de discuter de l'ensemble. Lorsqu'une publication a été réalisée ou est en cours de soumission concernant ce travail, une présentation de celle-ci sera également insérée.



Chapitre 7

Procédure de construction d'indices socio-économiques

Ce chapitre montre comment nous avons exploré la problématique de la mise au point d'indicateurs composites dans le contexte déjà présenté, ainsi que la procédure développée avec ses avantages et ses limites. Il décrit aussi le package R développé pour l'implémenter et quelques-uns des principaux résultats de l'application.

7.1 Contexte

Nous avons vu la quantité et la diversité d'informations disponibles, en particulier concernant les aspects démographiques et socio-économiques. Cette diversité, bien que présentant de nombreux avantages, peut être à l'origine de difficultés pour le chercheur s'intéressant aux inégalités sociales (dont on a rappelé l'importance sur la santé en section 1.1), dans la mesure où il doit alors déterminer comment mesurer le statut socio-économique.

Une solution possible est de n'utiliser qu'un petit nombre de variables sélectionnées de manière raisonnée (en fonction de la littérature notamment). La question du choix d'un indicateur plutôt qu'un autre est alors fréquemment posée. C'est néanmoins ce que font de nombreuses études, qui n'utilisent qu'une seule variable [4, 15, 61, 130], ou une petite sélection, comme indicateur socio-économique. Ces études peuvent alors explorer un par un les liens entre variables socio-économiques et événement de santé considéré, ou encore utiliser des modèles multivariés incluant toutes les variables socio-économiques retenues comme variables explicatives. Ces solutions présentent néanmoins des problèmes, liés notamment aux corrélations qui peuvent exister entre les variables.

Mais le concept de statut socio-économique est complexe et implique de nombreux aspects : emploi, revenu, éducation, logement, liens sociaux ... [99, 100, 103, 131–133]. Une alternative est alors d'utiliser toutes les données, mais de les synthétiser afin d'éviter les difficultés liées à la colinéarité et aux corrélations. Il s'agit alors de construire et d'utiliser un ou des indices socio-économiques composites (c'est par exemple le cas des indices développés par Pampalon *et al.* [104], qui réunissent un indice socio-économique « matériel » et un autre « social), dans notre cas à l'échelle de l'IRIS, qui synthétisent les différents aspects du statut socio-économique. Cette approche a conduit depuis la fin des années 1980 [95, 98] à la création de nombreux indices [81–83, 95–97, 99–102, 104, 105, 134–136].

Ceux-ci peuvent alors servir autant comme variables explicatives dans les modèles que directement comme indicateurs composites pour aider les acteurs locaux à déterminer des zones d'action publique.

Cependant, la plupart des indices existants posent différents problèmes. La sélection des variables qu'ils utilisent est principalement basée sur le choix arbitraire de variables reconnues dans la littérature comme étant de bons indicateurs du statut socio-économique. Cette sélection étant faite une fois pour toute, cela conduit à utiliser encore actuellement des indices parfois définis il y a plus de vingt ans, alors même que les définitions des variables ou leur importance vis-à-vis du statut socio-économique a pu évoluer.

De même, cette sélection de variables est adaptée pour le pays et le type de zone (rurale, urbaine) pour laquelle elle a été prévue, mais ne l'est pas forcément à d'autres contextes. La comparaison entre pays peut ainsi être rendue difficile à cause des différences de définitions (la notion de « classe sociale » utilisée dans certains indices peut fortement changer selon les pays, quand elle y est définie). Il en est de même pour les comparaisons au sein des pays mais dans des contextes différents (l'interprétation de la proportion de foyers sans voiture peut dépendre de l'infrastructure et de la disponibilité des réseaux de transports en commun locaux, ou selon que l'on est en contexte rural ou urbain). Ceci rend donc parfois la reproductibilité de ces indices compliquée.

Une fois les variables sélectionnées, la plupart des indices existants n'en utilise qu'un nombre limité (moins de dix), ce qui peut ne pas donner une représentation complète du statut socio-économique, ou limiter son utilisation par les acteurs locaux. Deux des indices les plus couramment utilisés, définis par Townsend [96] et Carstairs [97] n'utilisent par exemple que quatre variables choisies à partir de la littérature.

Indice de Townsend

L'indice de Townsend est défini comme la somme non pondérée des pourcentages *centrés-réduits* de foyers sans voiture et de ménages non propriétaires de leur logement, ajoutés aux log-pourcentages centrés-réduits du chômage et des logements surpeuplés.

$$I_{\text{Townsend}} = \log(\text{chômage}) + \log(\text{surpeuplement}) + \text{sans\_voiture} + \text{non\_proprio}$$

Indice de Carstairs

L'indice de Carstairs d'une zone est défini comme la somme non pondérée des pourcentages *centrés-réduits* de chômage, de population de classe sociale faible (selon la construction britannique des classes sociales), de foyers sans voiture et de surpeuplement des logements.

$$I_{\text{Carstairs}} = \text{chômage} + \text{classes\_sociales} + \text{surpeuplement} + \text{sans\_voiture}$$

Enfin, la méthode d'agrégation des variables retenues est souvent simple et consiste par exemple en des sommes (éventuellement pondérées) de quelques variables socio-économiques [82, 95, 97, 99, 100, 134].

On peut néanmoins préciser que quelques indices utilisent également des techniques d'analyse de données comme l'analyse en composantes principales (ACP, voir section 7.2.1) ou l'analyse en facteurs communs et spécifiques [82, 99, 101, 102, 104, 105], des techniques géostatistiques ou autres [81, 83, 135, 136].

Ces techniques plus élaborées sont cependant rarement utilisées dans le domaine de l'épidémiologie. Une explication possible à cela est que, comme nous l'évoquions en introduction générale, la mise au point et le développement d'une étude épidémiologique occasionne de nombreuses réflexions méthodologiques lors de sa conception et de sa mise en œuvre (par exemple pour éviter des biais de sélection ou de recrutement, définir précisément l'événement sanitaire d'intérêt, recueillir les données sur le terrain, obtenir les différentes autorisations requises comme celle de la CNIL, ...) et les analyses statistiques n'arrivent bien souvent qu'en bout de chaîne. Ceci implique bien souvent que les techniques utilisées sont choisies en fonction de ce qui est réalisé dans la littérature et des « traditions » du domaine, afin d'utiliser des méthodes bien connues (à la fois du chercheur et de ses interlocuteurs) plutôt que d'ajouter une réflexion et une difficulté supplémentaire dans la prise en compte d'une notion connue (ici le statut socio-économique).

> Pour ces raisons, nous avons choisi d'adopter une position intermédiaire, en réalisant une sélection, d'abord raisonnée puis sur des critères statistiques, de variables socio-économiques qui seront ensuite synthétisées en un indice. Pour éviter les limites précédemment énoncées des indices existants, nous avons souhaité créer une procédure de création d'indices qui soit flexible et facilement reproductible afin de permettre son utilisation dans des contextes multiples. Cet objectif nous a rapidement conduit à nous diriger vers les techniques d'analyse de données.

Une autre problématique réside dans la catégorisation des indices socio-économiques construits. Bien que ceux-ci soient souvent utilisés comme des indicateurs quantitatifs, il est fréquent également de les catégoriser à des fins de cartographie ou encore pour étudier une relation potentiellement non linéaire entre le statut socio-économique et la santé. Le plus souvent, une discrétisation par quantiles est utilisée mais ceci implique que les unités statistiques (ici les IRIS) sont regroupées uniquement sur un critère de taille de classes. Or, il paraîtrait plus cohérent de regrouper les unités statistiques sur des critères d'homogénéité des classes vis-à-vis du statut socio-économique afin d'obtenir des représentations graphiques plus fidèles. Ce constat nous a donc conduit à étudier les différentes méthodes de classification se basant sur les résultats de cette procédure.

> Enfin, dans la mesure où la procédure détaillée ici est plus complexe que celles habituellement utilisées dans le domaine de l'épidémiologie sociale et toujours dans un objectif de pouvoir reproduire cette procédure plus facilement, nous avons également développé un package pour le logiciel R implémentant celle-ci de manière à la rendre plus aisément utilisable.

7.2 Présentation de la procédure et du package R

7.2.1 Procédure de création d'indices socio-économiques

La procédure de création d'indices socio-économiques présentée ici est l'amélioration et l'extension d'une autre procédure précédemment développée par la même équipe [137] afin

de la rendre statistiquement plus cohérente et plus facilement généralisable. Étant donné la nature des variables socio-économiques utilisées (et détaillées dans le chapitre 5) et les objectifs de cette procédure, le choix a été fait d'utiliser les techniques d'analyse de données et en particulier l'analyse en composantes principales (ACP, dont le principe est rappelé en annexe A.1.1).

Nous avons choisi cette technique pour plusieurs raisons :

- Tout d'abord, l'objectif étant ici d'avoir une approche « guidée par les données », l'utilisation d'une technique d'analyse de données est apparue comme naturelle. Le choix de l'ACP résulte alors de la nature des données, toutes quantitatives, ainsi que du type d'indice attendu comme résultat. En effet, on souhaite ici avoir un indice quantitatif, le domaine de la santé publique cherchant souvent à mettre en évidence des gradients de risque et préférant disposer en conséquence de variables quantitatives ou qualitatives ordinales.
- De plus, l'ACP est une technique qui, bien que n'étant pas encore employée de manière intensive dans ce domaine, est tout de même déjà connue. Ce point nous permet donc de nous assurer que la procédure sera relativement facile à décrire, à mettre en œuvre et à interpréter pour les acteurs de ce domaine d'application, y compris non-statisticiens.

La procédure de création d'indices socio-économiques est divisée en trois étapes (voir figure 7.1) :

1) Réduction des groupes de variables redondantes

Ainsi que nous l'avons détaillé dans la première partie (section 5.1.2), des variables « redondantes » représentant la même notion peuvent être incluses dans la sélection initiale de variables. Or, ces variables sont d'une part très proches dans leurs définitions, mais également fortement corrélées. Par conséquent, la première étape de la procédure consiste à choisir, pour chacun des groupes de variables redondantes une unique variable représentant le mieux le groupe.

En effet, si dans le cas de l'illustration présentée ici (avec le groupe des variables de chômage, par exemple) le choix de la « meilleure variable » peut sembler simple et intuitif, ce n'est en réalité pas toujours le cas car elle est spécifique à chaque agglomération. Ainsi, bien que le chômage total ait été sélectionné dans la plupart des applications, c'est par exemple le chômage de longue durée qui est apparu plus pertinent dans le cas de l'agglomération de Lille avec les données du recensement 1999. De plus, dans la mesure où la procédure développée ici a pour but d'être générale et reproductible, nous avons préféré maintenir cette étape qui peut s'avérer utile dans le cas de contextes différents de ceux testés ici.

Pour cela, une ACP est appliquée sur chaque groupe de variables concerné. Le premier facteur de cette analyse est alors une bonne synthèse de l'ensemble du groupe, corrélé fortement avec l'ensemble des variables du groupe (dans le cas contraire, la constitution des groupes de variables redondantes n'est peut-être pas bonne). Cependant, afin de faciliter l'interprétation finale seule la variable ayant la plus forte corrélation avec ce premier facteur est conservée.

Dans le cas où il n'y aurait pas de variables « redondantes » dans la sélection initiale, cette étape peut être omise.

2) Sélection des variables

Une ACP est effectuée sur les variables retenues (celles issues de la première étape et celles qui ne faisaient pas partie des groupes de variables redondantes) pour sélectionner les variables qui rentreront dans l'indice final. Pour cela, seules les variables ayant une contribution au premier facteur supérieure à la moyenne sont conservées, ce qui revient à ne garder que les variables les plus corrélées au premier facteur.

Étant donné la répartition des différentes variables socio-économiques en domaines (emploi, éducation, logement, ...), nous avons également étudié la possibilité d'utiliser une analyse factorielle multiple (AFM , dont le principe est rappelé en annexe A.1.2) à cette étape. Cependant, l'objectif ici est de sélectionner les variables expliquant le plus de contrastes socio-économiques au sein de la zone étudiée sans considérations d'égalité entre les groupes de variables.
Nous avons cependant comparé les résultats obtenus en utilisant une ACP ou une AFM à cette étape en étudiant à la fois les sélections de variables réalisées à cette étape, leurs corrélations avec l'indice créé et les valeurs des indices proprement dits. Nous avons ainsi pu constater que si les sélections faites par l'AFM comprenaient moins de variables (une quinzaine) que les sélections par ACP (une vingtaine de variables), celles-ci comptaient néanmoins dix variables communes (aux techniques et aux agglomérations étudiées) (voir table 7.1). De plus, les corrélations entre ces variables communes et l'indice créé étaient proches pour les deux techniques, et les tests de comparaison de moyenne et/ou de Wilcoxon pour échantillons appariés appliqués aux valeurs de l'indice étaient non significatifs pour toutes les villes.

En prenant en compte ces résultats et l'objectif précédemment exposé nous avons donc décidé de n'utiliser que l'ACP pour répondre à l'objectif initial.

3) Construction de l'indice final

Une dernière ACP est réalisée sur les variables sélectionnées à la deuxième étape. Ici, nous avons adapté l'utilisation des résultats de l'ACP au contexte d'application, pour lequel on ne souhaitait qu'un unique indice socio-économique composite. Ainsi, sous réserve que le premier axe de cette ACP puisse bien être interprété comme un « axe socio-économique », ce qui est attendu étant donné les variables de départ mais n'est pas certain *a priori*, on définit l'indice socio-économique comme étant le premier facteur réduit. On obtient alors un indice de moyenne 0 et d'écart type 1.

Dans la mesure où l'on souhaite ici un unique indice, les axes suivants de l'ACP ne sont pas utilisés. Notons par ailleurs que ces axes ne présentaient jamais d'interprétation claire dans les applications de la procédure ayant été effectuées.

Cette procédure part donc d'un ensemble de variables choisies suivant la littérature de la même manière que de nombreux autres indices [11, 82, 104, 105, 134, 138] mais sélectionne ensuite parmi cet ensemble les variables qui rentreront dans la composition de l'indice final.

TABLE 7.1 – Variables sélectionnées à l'étape 2 de la procédure par AFM ou ACP, par agglomération, pour les données 1999

	ACP			**AFM**		
	Marseille	**Lyon**	**Lille**	**Marseille**	**Lyon**	**Lille**
Variables communes aux deux techniques	Population étrangère Immigrés étrangers Familles monoparentales Sans diplômes Bac - Brevet Bac +2 Emplois stables Revenu médian					
Variables communes à toutes les agglomérations, par technique	Emplois précaires Non salariés Non propriétaires de leur logement Plus d'une personne par pièce Nombre moyen de personnes par pièce Sans voiture Deux voitures ou plus			Moins de 25 ans		
Autres variables	Chômage total Cadres - Maisons Immeubles - Garages -	Chômage total Cadres Ouvriers - - HLM - -	Chômage > 1 an - - Maisons Immeubles HLM Garages Moins de 25 ans	Chômage total - Non salariés Cadres - Non proprétaires - - - - -	Chômage total Emplois précaires Non salariés Cadres Ouvriers - HLM > 1 pers./pièce - - Etudes supérieures	Chômage > 1 an Emplois précaires - - - - - > 1 pers./pièce Nb moyen de pers./pièce Sans voiture -
Nombre de variables sélectionnées	20	19	21	13	17	14

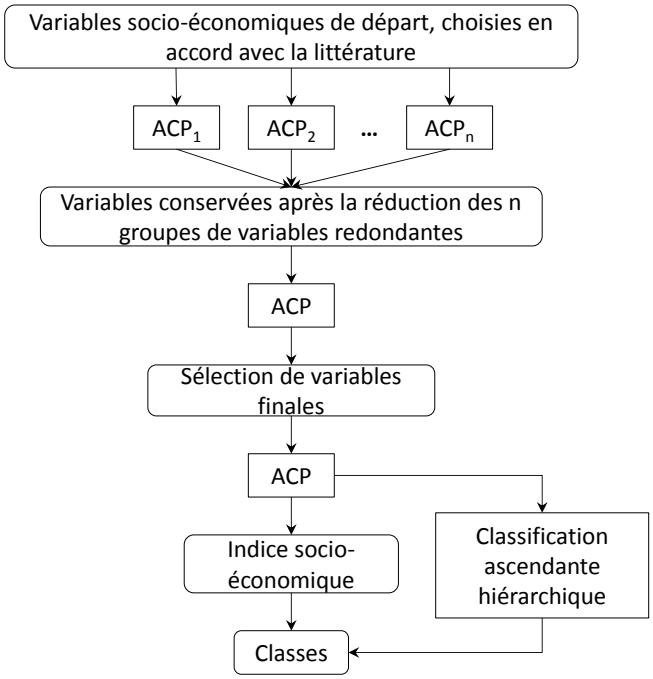


FIGURE 7.1 – Schéma de la procédure de création d'indices socio-économiques.

Cette sélection basée sur des critères statistiques est dirigée par les données, ce qui permet de s'affranchir d'une partie de la subjectivité pouvant influencer le choix des variables mais aussi de ne pas être dépendant d'une sélection fixe de variables, est l'une des originalités de cette procédure.

Le nombre de variables sélectionnées, qui n'est pas défini *a priori*, est également en général plus important que pour les autres indices existant. Ceci permet d'avoir une interprétation plus fine des résultats et de fournir davantage d'informations pour les acteurs locaux et les décideurs : une fois certaines zones identifiées grâce à l'indice socio-économique, rien n'empêche en effet de retourner au détail des variables pour expliciter les spécificités des zones ou déterminer les variables qui pourraient être un levier d'action.

Notons cependant que cette procédure n'a ici été appliquée et testée que pour créer des indices contextuels et non individuels. Par conséquent, les indices obtenus ne reflètent qu'un statut socio-économique des IRIS (dont on a discuté de l'homogénéité en section 3.1.1) et chercher à les interpréter à l'échelle individuelle peut être risqué étant donné l'existence du biais écologique. Cependant, il paraît parfaitement possible d'utiliser directement la procédure sur des données individuelles afin d'obtenir un indice socio-économique individuel.

Parmi les limites de cette procédure, l'une d'elle est partagée par tous les indices de ce type. Ainsi, il s'agit d'indicateurs composites généralement sans unité ni valeur de référence « intuitive », ce qui conduit à avoir une interprétation « relative » (à la zone sur laquelle a été construite l'indice) et parfois compliquée dans les applications, que ce soit lorsqu'ils sont utilisés seuls ou dans des modèles (l'interprétation de l'augmentation d'une unité de l'indice

socio-économique n'étant pas forcément simple à exprimer). Par ailleurs, du point de vue de l'application en santé publique, donner simplement un indice socio-économique peut certes permettre d'identifier les IRIS plus ou moins défavorisés, mais ne permet pas de connaître les déterminants plus précis de cette défaveur (d'où l'utilité évoquée précédemment de revenir aux variables, et donc que celles-ci soient en nombre relativement important).

Toujours vis-à-vis du domaine d'application, cette procédure nécessite (par conception) davantage de données et est également plus complexe à mettre en œuvre que la plupart des indices déjà existants. Cela nous paraît néanmoins être un prix à payer pour avoir une analyse plus en profondeur des liens entre les différentes variables et des contrastes socio-économiques entre les aires géographiques étudiées. En donnant une procédure plutôt qu'un indice « tout fait », nous permettons également de définir des indices dans des contextes, des zones et des périodes différentes mais construits suivant le même principe.

7.2.2 Classifications

À la suite de cette procédure de création, on obtient donc un indice socio-économique quantitatif synthétisant les informations contenues dans les différentes variables sélectionnées. Cependant, comme nous l'avons évoqué précédemment, dans le domaine d'application de la santé publique, l'inclusion d'une variable explicative dans des modèles explicatifs a souvent pour but de déterminer si un gradient de risque existe. Utiliser une variable quantitative dans ces modèles impliquera souvent que l'on fait une hypothèse de linéarité. Or, ce n'est pas le cas de l'application que l'on effectue pour ce travail, où l'on fait plutôt l'hypothèse que le risque de mortalité infantile est beaucoup plus élevé dans les catégories de statut socio-économique très faible par rapport aux catégories plus aisées.

Par ailleurs, l'une des clés de communication (aussi bien scientifique que face à des décideurs) dans les études « écologiques » de type géographique réside bien souvent dans la **mise à disposition de représentations cartographiques**.

Par conséquent, il est nécessaire d'étudier la discrétisation de l'indice quantitatif obtenu afin de répondre à ces objectifs du domaine d'application.

Comme nous l'avons exposé lors de la présentation du contexte, l'utilisation des quantiles pour réaliser des classes est extrêmement courante dans ce domaine d'application mais peu satisfaisante. En effet, si une discrétisation par quantiles est très facile à mettre en œuvre et offre l'avantage de donner des classes d'effectifs égaux, cette dernière propriété implique également que les classes constituées ne seront pas forcément optimales en termes d'homogénéité.

Des méthodes d'analyse de données comme la classification ascendante hiérarchique (CAH, dont le principe est rappelé en annexe A.1.3) permettent quant à elles de créer des classes basées sur la ressemblance des individus statistiques vis-à-vis des variables. Si effectuer une CAH sur les facteurs d'une analyse factorielle est très courant dans un cadre d'analyse de données, ceci l'est beaucoup moins dans le domaine de l'étude des inégalités sociales de santé où l'ACP n'est pas fréquente et où la CAH n'est presque jamais utilisée à notre connaissance. C'est pourquoi nous avons souhaité employer cette technique ici et cherché à montrer son intérêt dans ce domaine d'application.

Une autre raison qui nous a conduit à l'utilisation de la CAH est de pouvoir mener une **réflexion sur le nombre de classes à considérer**. Généralement, cinq classes sont utilisées en épidémiologie sociale, à la fois pour permettre de mettre en évidence des gradients de risque plus « fins » qu'avec un nombre de classes plus réduit, et afin d'avoir des représentations cartographiques lisibles. Mais ce nombre pourrait ne pas être adapté d'un point de vue statistique sur la pertinence de ces classes. À l'aide de la CAH, nous pourrons donc tenir compte de ces considérations et estimer quels compromis sont possibles pour obtenir un nombre de classes pertinent à la fois d'un point de vue statistique et dans le domaine d'application.

Cependant, la CAH est une technique multidimensionnelle qui peut utiliser plusieurs facteurs d'une ACP. Or, bien qu'il soit également possible de produire directement des catégories socio-économiques à partir de l'ACP finale de la procédure de création de l'indice, le but est ici de construire des classes à partir d'un indice unidimensionnel. La **CAH est ainsi utilisée ici comme « référence » pour déterminer des seuils optimaux** permettant de construire des classes unidimensionnelles aussi proches que possibles de celles obtenues par CAH.

Deux cas de figure sont alors possibles (en fonction du nombre de classes choisi) :

- Soit les classes construites par CAH sont distribuées sur plusieurs axes différents en plus du premier. Dans ce cas il est impossible de déterminer des seuils optimaux uniquement le long du premier axe capables d'approximer correctement les classes par CAH. L'indice ne peut alors être utilisé seul pour construire ce nombre de classes et ce dernier est réduit.
- Soit les classes sont distribuées le long du premier axe de l'ACP (donc de l'indice socio-économique). Dans ce cas, les seuils optimaux sont déterminés en utilisant un algorithme itératif simple : on fait parcourir les différentes valeurs possibles de l'indice aux seuils. À chaque étape, un taux de concordance entre la classification par seuils et celle par CAH est calculé (le pourcentage d'IRIS classés de manière similaire dans les deux méthodes). Seuls les seuils permettant d'obtenir le meilleur taux de concordance entre les deux classifications sont conservés.

Il est ainsi possible d'obtenir des classes proches de celles obtenues par CAH, mais n'utilisant que les valeurs de l'indice socio-économique pour être définies. L'algorithme qui a été utilisé n'est cependant pas optimal et demande rapidement un temps de calcul important avec l'augmentation du nombre d'individus statistiques et/ou de classes.

Nous avons implémenté cet algorithme (dans R) puis étudié les concordances entre les discrétisations obtenues par celui-ci, par CAH et par quantiles dans le cas de différents nombres de classes (entre 3 et 5, dans nos applications). Pour cela, nous avons utilisé le taux de concordance cité précédemment, mais également des indicateurs comme le κ de Cohen [139] ou l'indice de Rand [140]. Ces deux indices permettent également de mesurer l'accord ou le désaccord entre différentes classifications.

Coefficient κ de Cohen

Soit :

- P_o la concordance observée entre les deux classifications (i.e. le taux de concor-

dance),
- P_e la concordance aléatoire calculée à partir des marges de la table de contingence (i.e. la valeur espérée sous l'hypothèse d'indépendance des classifications),

alors le coefficient κ est défini comme

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

À partir de ce coefficient il est possible d'avoir une évaluation du degré « d'accord » des deux méthodes de classification, celui-ci étant d'autant plus important que κ se rapproche de 1.

Indice de Rand

L'indice de Rand compare les similarités et dissimilarités entre deux classifications en prenant en compte à la fois les unités classées de la même manière et les unités séparées dans les deux classifications.

Ainsi, si l'on dispose de :
- n points $x_1, \ldots, x_n$ et
- deux classifications $Y = \{Y_1, \ldots, Y_{k_1}\}$ et $Y' = \{Y'_1, \ldots, Y'_{k_2}\}$,

l'indice de Rand de ces classifications est

$$c(Y, Y') = \frac{\sum_{i<j}^{n} \gamma_{i,j}}{\binom{n}{2}}$$

où

$$\gamma_{i,j} = \begin{cases} 1 & \text{si } x_i \text{ et } x_j \text{ appartiennent à la même classe dans les deux classifications} \\ 1 & \text{si } x_i \text{ et } x_j \text{ sont séparés dans les deux classifications} \\ 0 & \text{sinon } (x_i \text{ et } x_j \text{ sont regroupés dans une classification et séparés dans l'autre}) \end{cases}$$

Les principaux résultats de ces comparaisons sont exposés dans la section suivante.

Ce type de classification présente néanmoins différentes limites. Tout d'abord, la CAH crée des classes homogènes dans leur composition, mais sans critère de taille. Ceci peut donc conduire à obtenir des classes aux effectifs très différents, pouvant impliquer par la suite des problèmes de précision d'estimations lorsque l'on souhaite lier ces classes à des événements de santé dans les applications.

L'algorithme qui a été utilisé partage ces limites. Il devra être optimisé car il demande rapidement un temps de calcul important. En effet, dans son implémentation actuelle, nous avons fait appel à des boucles imbriquées dont la quantité augmente avec le nombre de seuils à déterminer. L'étendue des valeurs à parcourir pour chaque boucle étant de plus lié au nombre d'unités étudiées, cela implique un temps de calcul proportionnel à la fois au nombre de classes et au nombre d'individus. Par ailleurs, si l'on ne s'est pas assuré que les classes sont bien réparties le long du premier axe, l'algorithme actuel peut ne pas

converger en un temps raisonnable. Malgré l'intérêt de cet algorithme, sa complexité de mise en œuvre et son temps d'exécution n'en font donc à l'heure actuelle pas une alternative raisonnable à la CAH (ou aux quantiles).

7.2.3 Package R

Pour finir, la procédure de création présentée ci-dessus a été implémentée dans un package R[141] nommé SesIndexCreator. En effet, ainsi que nous l'avons exposé en introduction, l'un des objectifs de cette procédure est d'être flexible et facilement reproductible dans différents contextes. Elle peut ainsi être utilisée sur différentes zones géographiques, à différentes échelles ou encore sur différentes périodes. Par exemple, la nouvelle définition du recensement de la population impliquant que les différentes statistiques seront mises à jour tous les ans, une utilisation de la procédure « en routine » afin d'examiner l'évolution du statut socio-économique de certaines zones dans le temps pourrait être envisagée.

> C'est pour ces raisons que nous avons développé ce package, avec pour but de simplifier l'utilisation de la procédure et l'interprétation de ses résultats afin de la rendre aussi accessible que possible et de permettre son utilisation sur d'autres zones ou périodes d'étude.

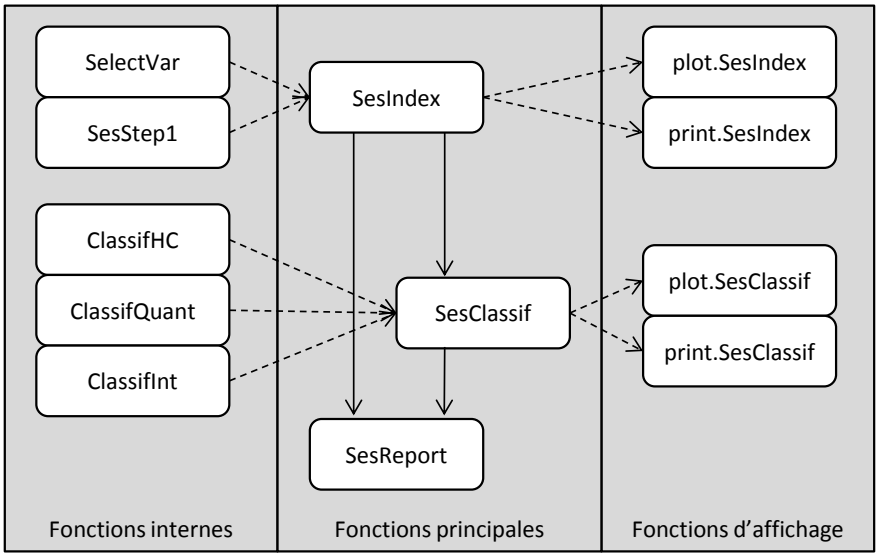


FIGURE 7.2 – Organisation des fonctions du package SesIndexCreatoR.

Ce package dépend en particulier du package FactoMineR [142], qui implémente de nombreuses fonctions d'analyse de données. SesIndexCreatoR, que nous avons déposé sur le CRAN [143], est composé d'une dizaine de fonctions dont trois fonctions principales permettant d'appliquer la procédure de création, de créer des catégories à partir de celle-ci puis de générer des rapports automatiques. Ces différentes fonctions sont détaillées dans l'article présenté dans la section 7.5.

Si le cœur de ce package permet d'appliquer la procédure de manière simplifiée, il demeure limité en taille et fonctionnalités à l'heure actuelle. Des améliorations pour rendre

son utilisation plus aisée, mais aussi pour ajouter différentes fonctionnalités sont ainsi envisagées. Ces fonctionnalités pourraient aussi bien avoir un intérêt méthodologique, comme l'ajout d'autres méthodes de classification, que pratique, comme l'intégration d'outils de représentations cartographiques ou d'autres aides à l'interprétation.

7.3 Résumé des principaux résultats

La procédure décrite dans la section précédente a été appliquée sur chacune des quatre agglomérations de l'étude avec comme variables de départ les données socio-économiques présentées en première partie (sections 5.1.3 et 5.1.4), d'abord avec les données issues du recensement de 1999 puis avec celles du recensement de 2006.

La procédure a également été appliquée en une analyse « globale » sur les quatre agglomérations simultanément. En effet, l'un des inconvénients de la procédure présentée ici est que l'indice est relatif à la zone sur laquelle il a été créé, ce qui empêche de comparer les valeurs de l'indice socio-économique des IRIS de deux agglomérations s'il n'a pas été créé sur ces deux agglomérations en même temps. L'analyse « globale », en construisant l'indice sur l'ensemble des quatre agglomérations en même temps, permettra donc de comparer les agglomérations entre elles. Le choix d'un indice spécifique à une agglomération ou d'un indice global dépend ainsi directement de l'objectif fixé : étudier les particularités et l'organisation des IRIS au sein d'une unique agglomération, ou comparer les positions respectives des agglomérations entre elles. Nous présenterons dans cette section les résultats de ces deux approches en parallèle, mais par la suite les indices spécifiques à chaque agglomération seront systématiquement utilisés.

Les IRIS d'activité et divers (voir section 3.1.2) ayant révélé des caractéristiques socio-économiques très particulières (et pouvant être considérées parfois comme des observations aberrantes : par exemple des IRIS avec 100% de retraités et de ménages composés d'une seule personne) et comptant par définition peu de résidents, ceux-ci ont été considérés comme des individus illustratifs dans les procédures afin qu'ils ne perturbent pas les résultats.

Dans cette section, nous présenterons une brève synthèse des principaux résultats obtenus avec les données 2006 et quelques résultats des données 1999 (détaillés dans l'article exposé en section 7.4).

Pour les données du recensement 2006 utilisées ici, trois groupes de variables « redondantes » ont été définis :

- le groupe des variables de population active : population active totale, hommes et femmes
- le groupe des variables de chômage : chômage total, des 15-24 ans, des plus de 55 ans, hommes et femmes
- le groupe des variables de mobilité : ménages ayant déménagé depuis moins de 2 ans, moins de 5 ans, plus de 10 ans, et mobilité professionnelle

Pour chacune des cinq analyses, les variables sélectionnées à la première étape de la procédure comme meilleures représentantes de leur groupe sont les mêmes. Il s'agit de la population active totale, du chômage total et des ménages ayant déménagé depuis moins de 5 ans.

TABLE 7.2 – Variables sélectionnées à l'étape 2 de la procédure, par agglomération

	Paris Petite Couronne	**Aix-Marseille**	**Grand Lyon**	**Lille Métropole**	**Globale**
Variables communes à toutes les analyses	Population étrangère Immigrés étrangers Familles monoparentales Sans diplôme Bac +2 Chômage total Non propriétaires de leur logement Revenu médian				
Variables communes aux agglomérations de province	- - - - - - -	Professions intermédiaires Emplois stables Maisons Immeubles Garages Surface de plus de 100m^2 Sans voiture Deux voitures ou plus			Professions intermédiaires Emplois stables - - Garages Surface de plus de 100m^2 Sans voitur Deux voitures ou plus
Autres variables	Moins de 25 ans Ménages seuls - - Cadres Employés Ouvriers Non salariés CAP-BEP Études supérieures HLM	Moins de 25 ans - - - Cadres - - Non salariés - - -	- - - Artisans Cadres - Ouvriers Non salariés - - HLM	Moins de 25 ans - Emplois précaires - - - - - - - HLM	Moins de 25 ans - - - Cadres Employés Ouvriers Non salariés - - HLM
Nombre de variables sélectionnées	17	19	21	20	20

La deuxième étape a donc été appliquée sur les mêmes variables pour les quatre agglomérations et l'analyse globale. Cette étape a conduit à sélectionner les variables finales à inclure dans l'indice socio-économique. Ces sélections sont résumées dans le tableau 7.2.

La procédure a ainsi sélectionné entre 17 et 21 variables parmi les 41 variables restantes après l'étape 1. On peut noter que parmi ces variables, 8 sont communes à toutes les analyses (soit 38 à 47% des variables sélectionnées). Si l'on ne considère que les agglomérations de province, l'agglomération parisienne ayant des caractéristiques très particulières, ce sont 16 variables qui sont communes aux trois zones (soit 76 à 84% des variables sélectionnées). Ceci semble montrer d'une part que la procédure est robuste, et d'autre part que malgré les différences entre agglomérations il est possible d'expliquer une importante proportion de la variabilité socio-économique avec les mêmes variables communes. Les spécificités de chaque agglomération peuvent également être étudiées en observant les variables propres à chacune dans la sélection faite, mais aussi en étudiant la contribution des différentes variables à l'indice.

TABLE 7.3 – Pourcentage de variance expliquée par les premiers facteurs de l'ACP finale, par agglomération

	Paris	**Marseille**	**Lyon**	**Lille**	**Globale**
1er facteur	59,16%	59,44%	54,20%	62,78%	44,80%
2e facteur	13,72%	13,36%	19,97%	12,77%	19,49%
3e facteur	8,53%	8,14%	10,07%	6,47%	10,17%
4e facteur	5,80%	3,48%	3,70%	4,38%	5,74%
5e facteur	3,20%	3,25%	2,44%	3,20%	4,01%

Pour des raisons de place, seul le nom des villes principales des agglomérations a été indiqué.

La dernière étape permet de créer l'indice proprement dit. Dans chaque cas, le premier facteur de l'ACP finale explique une proportion très importante de la variance totale (voir table 7.3). On observe par ailleurs systématiquement, à l'exception de l'agglomération de Marseille, que le premier facteur est corrélé positivement à des variables de « défaveur » (chômage, familles monoparentales, HLM, sans diplôme, ...) et corrélé négativement à des variables de « faveur » (revenu médian, bac +2, cadres, ...) (voir table 7.4 et figure 7.3). Dans le cas de Marseille, les corrélations sont à l'opposé de celles des autres agglomérations (premier facteur corrélé négativement aux variables de « défaveur ») mais, les axes factoriels étant définis au sens près, on transformera les résultats obtenus de manière à ce que l'orientation du premier axe soit cohérente avec les autres agglomérations.

Pour chaque agglomération, les experts locaux associés au projet ont pu confirmer que l'indice socio-économique construit était globalement en adéquation avec leur connaissance du terrain (voir figure 7.4). L'étude des corrélations entre les indices construits via notre procédure et les deux indices de Townsend et Carstairs (qui sont parmi les plus connus et utilisés dans la littérature) n'était pas possible avec les données du recensement 2006 (la variable de surpeuplement n'étant pas disponible). On observe cependant des corrélations importantes (supérieures à 0,9) entre ces indices et ceux construits avec notre procédure appliquée aux données du recensement 1999, ce qui semble confirmer que la notion de

statut socio-économique mesurée par nos indices est proche de celle mesurée par ces deux autres indices.

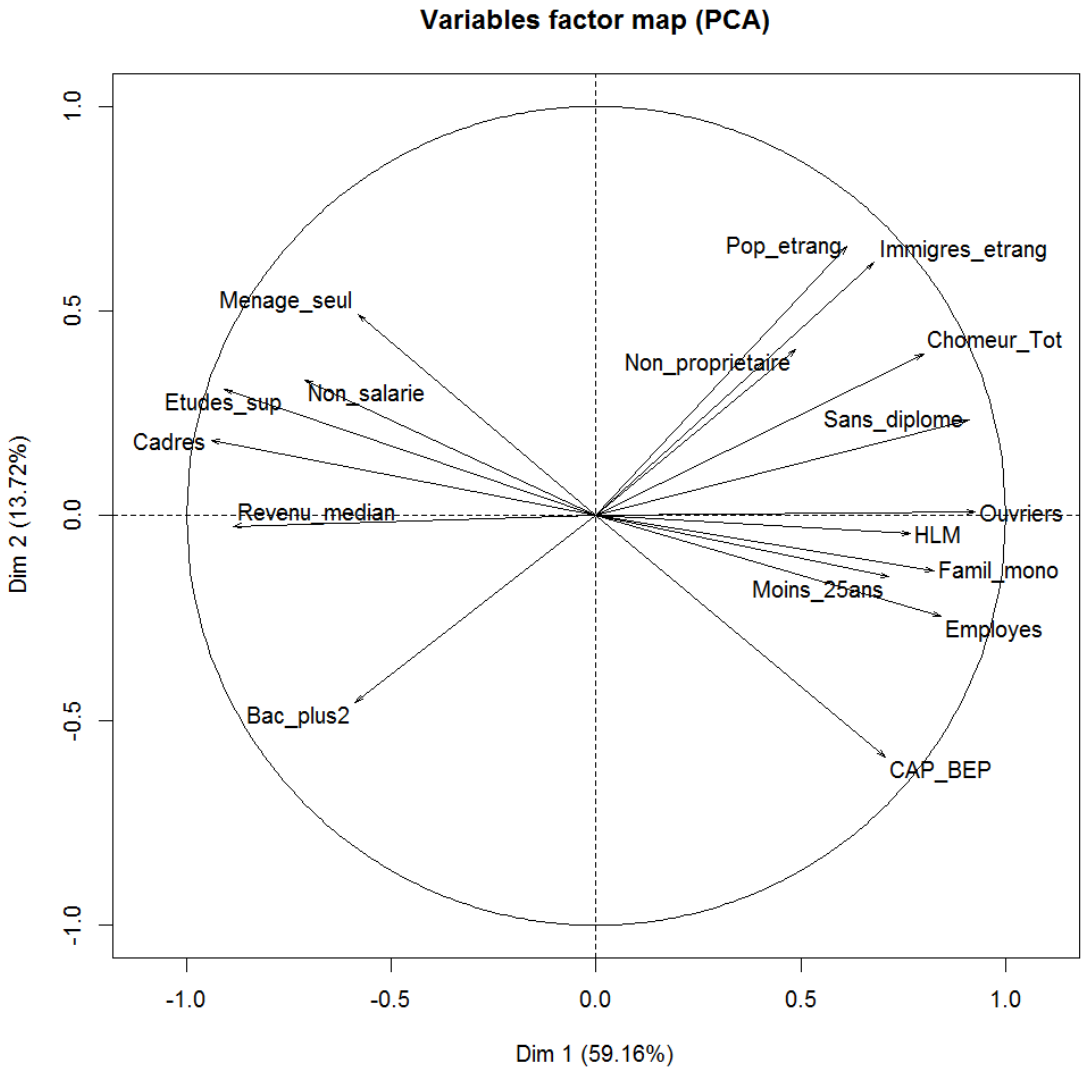


FIGURE 7.3 – Cercle des corrélations de l'ACP finale dans le cas de l'agglomération de Paris-Petite Couronne.

TABLE 7.4 – Corrélations et contributions des variables au premier facteur, par agglomération

Variables	Paris		Marseille[a]		Lyon		Lille		Globale	
	Coord[b]	Ctr[c]	Coord[b]	Ctr[c]	Coord[b]	Ctr[c]	Coord[b]	Ctr[c]	Coord[b]	Ctr[c]
Population étrangère	0,61	3,71	0,8	5,35	0,86	6,56	0,86	6,15	0,66	4,93
Immigrés étrangers	0,68	4,58	0,84	5,98	0,88	6,86	0,86	6,26	0,7	5,49
Familles monoparentales	0,82	6,74	0,73	4,53	0,67	3,95	0,67	3,77	0,75	6,22
Sans Diplôme	0,91	8,25	0,82	5,6	0,81	5,83	0,72	4,37	0,86	8,34
Bac +2	−0,59	3,44	−0,79	5,24	−0,78	5,28	−0,79	5,23	−0,74	6,16
Chômage total	0,8	6,34	0,91	6,92	0,9	7,13	0,9	6,85	0,86	8,17
Non propriétaires de leur logement	0,49	2,36	0,87	6,3	0,81	5,78	0,87	6,39	0,71	5,63
Revenu médian	−0,88	7,78	−0,89	6,61	−0,89	6,94	−0,84	5,85	−0,79	7,05
Professions intermédiaires	-	-	−0,78	5,06	−0,68	4,08	−0,75	4,73	−0,62	4,24
Emplois stables	-	-	−0,73	4,54	−0,62	3,41	−0,88	6,47	−0,56	3,5
Surface de plus de 100m^2	-	-	−0,74	4,63	−0,72	4,52	−0,76	4,83	−0,56	3,53
Sans voiture	-	-	0,82	5,71	0,63	3,45	0,89	6,64	0,42	1,96
Deux voitures ou plus	-	-	−0,79	5,19	−0,66	3,83	−0,86	6,22	−0,47	2,51
Garages	-	-	−0,86	6,19	−0,71	4,38	−0,88	6,55	−0,56	3,51
Maisons	-	-	−0,7	4,16	−0,56	2,74	−0,7	4,09	-	-
Immeubles	-	-	0,69	4,03	0,56	2,75	0,69	4,04	-	-
Moins de 25 ans	0,72	5,08	0,6	3,07	-	-	0,69	3,95	0,56	3,49
Ménages seuls	−0,58	3,37	-	-	-	-	-	-	-	-
Emplois précaires	-	-	-	-	-	-	0,62	3,24	-	-
Artisans	-	-	-	-	−0,61	3,24	-	-	-	-
Cadres	−0,94	8,77	−0,67	3,76	−0,69	4,14	-	-	−0,67	5,06
Employés	0,84	7,04	-	-	-	-	-	-	0,6	3,96
Ouvriers	0,92	8,5	-	-	0,75	4,94	-	-	0,74	6,1
Non salariés	−0,71	5,04	−0,63	3,37	−0,69	4,24	-	-	−0,59	3,9
CAP-BEP	0,7	4,93	-	-	-	-	-	-	-	-
Études supérieures	−0,91	8,21	-	-	-	-	-	-	-	-
HLM	0,77	5,84	-	-	0,82	5,96	0,72	4,37	0,75	6,26

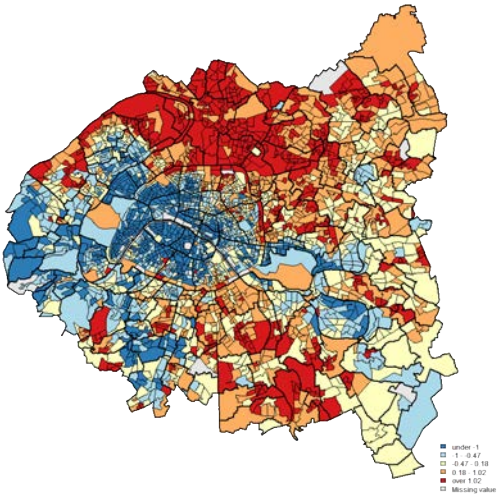
Pour des raisons de place, seul le nom des villes principales des agglomérations a été indiqué.

[a]Les résultats présentés ici pour l'agglomération de Marseille considèrent l'axe comme allant dans le même sens que pour les autres agglomérations (corrélation positive avec les variables de « défaveur »)

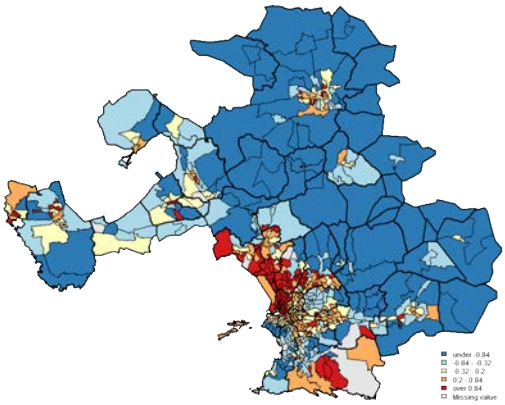
[b]Coordonnée de la variable sur le premier axe, i.e. coefficient de corrélation entre la variable et l'indice socio-économique

[c]Contribution de la variable au premier axe (%)

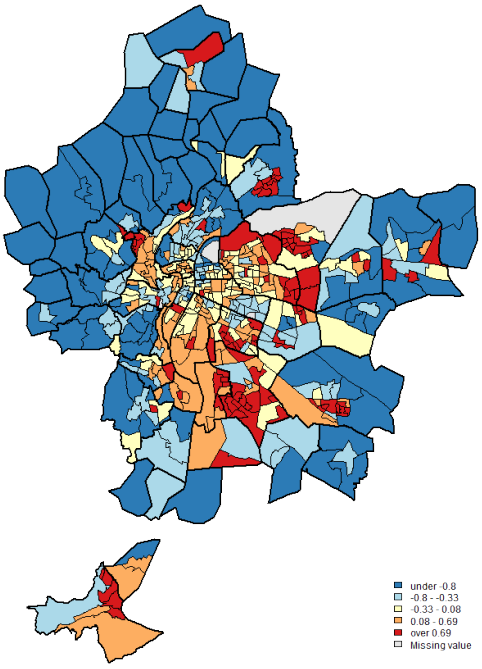
Italique : variables communes à toutes les analyses



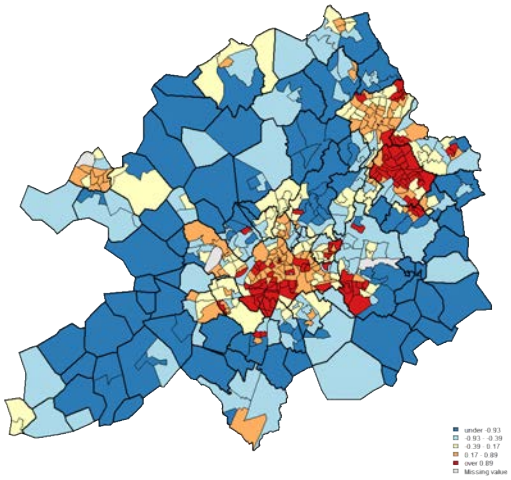
(a) Paris-Petite Couronne.



(b) Aix-Marseille.



(c) Grand Lyon.



(d) Lille Métropole.

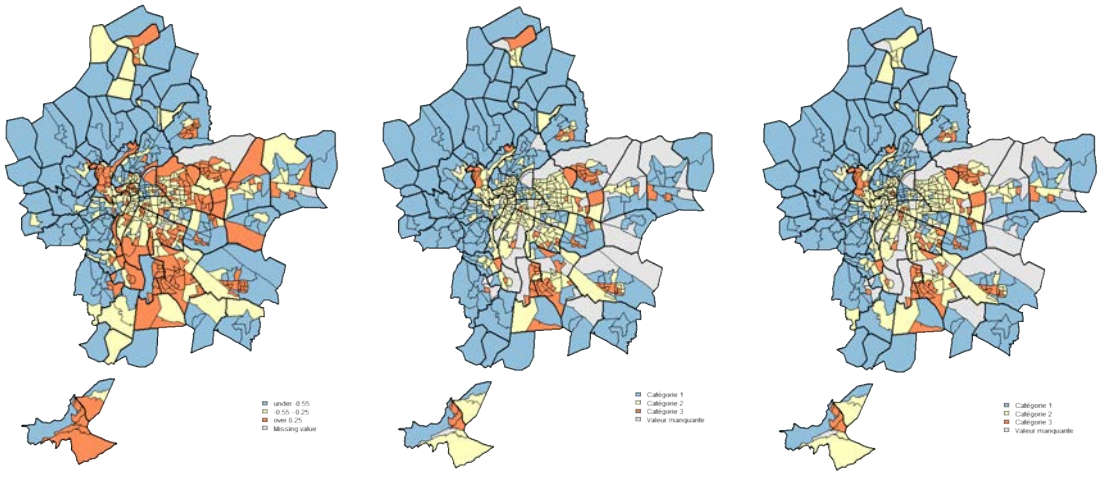
FIGURE 7.4 – Cartes de l'indice socio-économique sur les données 2006, en quintiles.

N.B. : les échelles géographiques sont différentes entre les cartes.

Concernant les méthodes de classification, nous avons principalement appliqué et étudié celles-ci sur les données 1999. Ceci nous a permis de nous rendre compte que dans notre cas, pour toutes les agglomérations étudiées, les cinq classes utilisées habituellement ne sont pas les plus pertinentes si l'on souhaite utiliser un indice unidimensionnel, et qu'il n'y a qu'une très faible concordance (environ 50%) entre les classes créées à l'aide de quintiles et celles créées à l'aide de la CAH. Ceci est en effet dû au fait que la CAH emploie dans ce cas les deux ou trois premiers facteurs de l'ACP servant à créer l'indice pour constituer ses classes. En réduisant le nombre de classes jusqu'à trois, il devient possible de créer des classes constituées uniquement le long du premier axe. La concordance entre la CAH et les quantiles demeure néanmoins modérément élevée (environ 70%) dans nos illustrations (voir table 7.5). La figure 7.5 illustre également les différences entre ces méthodes de classification dans le cas du Grand Lyon.

TABLE 7.5 – Taux de concordance entre les différentes méthodes de classification, données 1999

	CAH(5) vs quintiles	**CAH(3) vs tiertiles**	**CAH(3) vs seuils**	**Tiertiles vs seuils**
Globale	63%	71%	97%	72%
Aix-Marseille	48%	69%	97%	67%
Grand Lyon	48%	74%	93%	78%
Lille Métropole	41%	78%	98%	79%



(a) Par quantiles. (b) Par CAH. (c) Par seuils optimaux.

FIGURE 7.5 – Cartes de l'indice socio-économique sur les données 1999 avec différentes classifications, agglomération du Grand Lyon.

7.4 Article : A Statistical Procedure to Create a Neighborhood Socioeconomic Index for Health Inequalities Analysis

Un article présentant la procédure de création d'indices socio-économiques et la comparaison entre méthodes de classification précédemment exposées a été publié dans l'*International Journal for Equity in Health* en mars 2013. Celui-ci est présenté en annexe B.1.

Cet article présente également sous la même forme que dans la section précédente les résultats des analyses sur les données du recensement 1999. Dans la mesure où ces derniers sont très similaires à ceux présentés ici, nous ne les détaillerons pas davantage dans cette section.

De même, nous présentons dans cet article le détail des résultats des corrélations entre notre indice et les indices de Townsend et Carstairs et l'ensemble des résultats des différentes méthodes de classifications étudiées (ces résultats ont également été rappelés dans la section précédente) à l'aide du pourcentage de concordance, du coefficient κ et de l'indice de Rand.

Nous concluons cet article par une discussion sur la procédure de création d'indices socio-économiques que nous présentons, similaire à celle qui sera faite dans la section 7.7 de ce manuscrit.

7.5 Article : SesIndexCreatoR : An R Package for Socioeconomic Indices Computation and Visualization

L'article présenté en annexe B.2 est soumis depuis mai 2013 au *Journal of Statistical Software*.

Il présente le package SesIndexCreatoR exposé dans les sections précédentes plus en détail, en particulier en donnant un exemple « fil rouge » permettant une application et une explication pas à pas de l'utilisation des différentes fonctions du package afin d'appliquer la procédure, de constituer des classes à partir de l'indice obtenu puis de générer un rapport automatique synthétisant les résultats.

7.6 Applications de la procédure dans d'autres travaux

La procédure développée ici a été illustrée à plusieurs reprises tout au long de cette thèse. Ainsi, outre les résultats et les articles exposés dans les sections précédentes et se concentrant sur l'aspect purement socio-économique, **l'utilité des indices issus de cette procédure a pu être testée dans d'autres contextes**.

Tout d'abord, au sein du projet Equit'Area, cette procédure a été mise en œuvre en collaboration avec les travaux de différents membres de l'équipe. Différents travaux (publiés ou soumis) ont ainsi étudié :

Les inégalités sociales de mortalité infantile à Lille, à l'aide de modèles bayésiens (voir « *An Ecological Study to Identify Census Blocks Supporting a Higher Burden of Disease : Infant Mortality in the Lille Metropolitan Area, France* » en annexe C.2).

Dans ce travail, l'indice socio-économique construit à l'aide de notre procédure sur l'agglomération de Lille a été utilisé comme variable explicative, soit en tant que variable quantitative soit discrétisé en 5 classes (ce travail ayant eu lieu avant que nous ayons mené à son terme notre réflexion sur le nombre de classes).

Nous avons également réalisé l'analyse bayésienne de ce travail en mettant en œuvre le modèle « BYM » de manière similaire à ce qui sera exposé section 10.3 (nous détaillerons davantage l'apport de cette thèse de ce point de vue en section 10.4).

On montre dans cet article une augmentation du risque de mortalité infantile dans les IRIS les plus défavorisés par rapport aux plus favorisés, que l'indice socio-économique soit utilisé comme une variable explicative quantitative ou discrétisé et utilisé comme une variable explicative qualitative. On montre également que les IRIS les plus défavorisés et ayant les plus importants risques relatifs (estimés) se concentrent essentiellement dans les communes de Lille, Tourcoing et Roubaix.

L'influence des inégalités sociales et de l'exposition au NO_2 sur la mortalité infantile à Lyon et Lille, à l'aide de modèles additifs généralisés (voir « *Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France* » en annexe C.3).

A nouveau, nous avons créé et interprété l'indice socio-économique utilisé comme variable explicative (avec le dioxyde d'azote) dans les modèles additifs généralisés appliqués dans ce travail.

Les modèles additifs généralisés (que nous n'avons pas utilisés dans cette thèse) ont été introduits par Hastie et Tibshirani au début des années 1990 mêlant les modèles linéaires généralisés et les régressions non-linéaires. Ils sont de la forme $g(\mathbb{E}[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m)$ où l'on spécifie une distribution (de la famille exponentielle) pour Y et une fonction de lien g. Les f_i sont des fonctions paramétriques, semi-paramétriques ou non-paramétriques qui seront alors estimées différemment. Ici, on ajoute au modèle de Poisson une fonction des coordonnées spatiales (semi-paramétrique) de lissage, ce qui permet de prendre en compte l'aspect spatial des données.

Ce travail met en évidence des motifs de pollution différents entre les agglomérations, mais aussi des relations différentes entre les variables explicatives et la mortalité infantile suivant les villes.

Les inégalités sociales d'exposition à la pollution de l'air à Paris, Lyon et Lille à l'aide de modèles additifs généralisés (voir « *Air quality and Social deprivation in four French metropolitan areas ? A spatio-temporal environmental inequality analysis conducted at a small geographical level* » en annexe C.4).

Ce travail s'inscrit dans un contexte de « justice environnementale », c'est-à-dire dans l'étude des relations entre les expositions environnementales et le statut socio-économique.

Les modèles additifs généralisés sont à nouveaux utilisés, dans ce cas pour chercher à expliquer la moyenne annuelle de NO_2 par l'indice socio-économique que nous avons

construit et interprété en prenant également en compte les effets spatiaux. Chaque agglomération est étudiée indépendamment pour les deux périodes temporelles.

On montre dans cet article que les catégories sociales les plus exposées à la pollution de l'air ne sont pas les mêmes suivant les agglomérations. On montre également des différences entre les périodes suivant les agglomérations.

L'influence des inégalités sociales et du bruit sur la mortalité infantile à Lyon, à l'aide de techniques de balayage spatial (voir « *An exploratory spatial analysis to assess the relationship between deprivation, noise and infant mortality* » en annexe C.5).

Le statut socio-économique utilisé comme variable explicative est à nouveau représenté par l'indice que nous avons construit ou différentes variables socio-économiques (revenu, niveau d'éducation, chômage, ...).

La technique utilisée est une méthode d'analyse spatiale appelée « *spatial scan statistics* » (que nous n'avons pas utilisé dans cette thèse). Cette méthode effectue un balayage de la zone géographique suivant une grille définie automatiquement à partir de la structure spatiale des données ou directement par l'utilisateur (dans le cas présent, les centroïdes des IRIS sont utilisés). Pour chaque point de la grille, différentes fenêtres circulaires autour de ce point sont créées, avec des rayons allant de zéro à une limite supérieure fixée (ici, 50% de la taille de la zone). On compare le taux de mortalité infantile dans chacune de ces fenêtres par rapport au taux attendu sous l'hypothèse d'une distribution aléatoire (on suppose que le nombre de cas dans un IRIS suit une loi de Poisson), et on identifie les groupes les plus vraisemblables d'excès de risque à l'aide d'un test du rapport de vraisemblance.

Il est également possible d'inclure des variables explicatives, dans ce cas le bruit, le statut socio-économique ou les deux (avec interaction).

On montre dans cet article qu'il existe un groupe d'IRIS avec un risque plus élevé de mortalité infantile au sud-est de l'agglomération de Lyon. Ce groupe disparaît (ou se réduit) lorsque l'on ajoute le bruit et le statut socio-économique comme variables explicatives, ce qui suggère que le bruit et le statut socio-économique expliquent l'excès de risque.

L'influence des inégalités sociales et des espaces verts sur la mortalité néonatale à Lyon, à l'aide de techniques de balayage spatial (voir « *Green space, social inequalities and neonatal mortality in France* » en annexe C.6).

Dans ce travail, nous avons aussi créé l'indice socio-économique utilisé comme variable explicative dans la méthode « *spatial scan statistics* », cette fois avec les espaces verts comme variable d'exposition environnementale.

Ce travail est, à notre connaissance, le premier étudiant les liens entre la mortalité infantile, les espaces verts et le statut socio-économique.

Dans tout ces cas, l'apport de cette thèse a été d'appliquer la procédure de création d'indices socio-économiques dans leurs contextes respectifs, et nous avons apporté une expertise sur l'utilisation et l'interprétation des indices en résultant.

La confrontation de la procédure avec ces différents travaux nous a également permis de la tester, de la valider et de l'améliorer. Ainsi, par exemple, alors que la procédure n'avait dans un premier temps été testée que sur les données du recensement 1999 (comme on peut le voir dans les articles B.1 et C.2), la mise en ligne des données du recensement 2006 et la nécessité d'utiliser celles-ci pour les travaux suivants nous a conduit d'abord à reconstruire l'ensemble des indicateurs socio-économiques (présentés section 5.1.4) à partir de ces données, puis à appliquer la procédure sur ceux-ci afin que l'indice puisse être utilisé dans les articles C.3 à C.6.

Ceci nous a conduit à effectuer de nombreuses comparaisons entre les résultats obtenus sur les données 1999 et 2006 afin d'évaluer la reproductibilité et la stabilité de la procédure. Nous avons ainsi comparé les différentes sélections de variables faites aux étapes 1 et 2, mais également les corrélations et les contributions de ces variables à l'indice final. Nous avons ainsi pu constater que quatre variables étaient communes à toutes les villes et aux deux périodes : les proportions de familles monoparentales, de personnes sans diplômes, de non propriétaires de leur logement et le revenu médian. De plus, pour toutes les agglomérations, 7 à 12 variables sélectionnées sont communes (en plus des quatre précédentes) entre les deux années. Les corrélations et contributions de ces variables à l'indice final sont par ailleurs très similaires d'une année sur l'autre, ce qui semble donc indiquer une bonne reproductibilité de la procédure.

Le travail décrit dans ce manuscrit a également permis de faire évoluer notre avis sur le nombre de classes à employer. Ainsi, si dans les premiers travaux publiés (article C.2) l'indice socio-économique était discrétisé en cinq classes, nos réflexions sur le nombre de classes à utiliser ont conduit à adopter par la suite des discrétisations en trois classes.

Cette réflexion a aussi alimenté un travail dans un contexte légèrement différent du projet Equit'Area et s'intéressant à l'influence du statut socio-économique et de l'exposition au NO_2 à long terme sur le lien entre l'exposition au NO_2 à court terme et la mortalité toute-cause des plus de 35 ans à Paris (voir « *Do neighborhood characteristics modify the relation between short-term exposure to nitrogen dioxide and all-cause mortality ? A time-stratified case-crossover study conducted in Paris* » en annexe C.7). En effet, si le fait que des variations à court terme de la pollution (autrement dit, des pics de pollution) augmentent la mortalité est désormais bien documenté, l'influence de l'exposition à plus long terme et du statut socio-économique sur cette relation est moins connue.
Ici, ce sont directement les catégories de défaveur issues de la CAH qui ont été utilisées afin d'obtenir des catégories d'IRIS aussi homogènes que possible, pour ensuite intégrer ces résultats au sein de modèles de séries temporelles et de régressions logistiques conditionnelles.

D'autres travaux ont également alimenté notre réflexion sur la procédure elle-même ou sa reproductibilité. Ainsi, la collaboration avec BruitParif sur le projet SURVOL [144], nous a permis d'illustrer l'utilité de la procédure dans le cadre de l'évaluation des inégalités sociales face à l'exposition aux bruits de trafic aérien à proximité des aéroports parisiens.

La collaboration avec d'autres membres du département EpiBiostats de l'EHESP (Sahar Bayat et Fei Gao) a également alimenté notre réflexion sur la manière d'utiliser cette procédure dans des contextes mêlant zones rurales et urbaines.

Ainsi, dans le cadre d'un projet étudiant l'insuffisance rénale chronique terminale en Bretagne, nous avons pu constater que l'application de cette procédure sur toute une région tendait à transformer l'indice construit en un indice distinguant les IRIS ruraux et urbains. Une solution qui a été adoptée est alors de distinguer les IRIS suivant différents classes d'urbanisation et d'appliquer la procédure séparément sur chacune de ces classes.

La ville de Rennes apparaît alors avec des valeurs d'indice importantes (ce qui serait interprété comme une forte défaveur dans d'autres contextes) tandis que les zones plus rurales obtiennent des valeurs faibles d'indice. Ces différences sont néanmoins liées a priori aux variables sélectionnées, qui n'ont pas la même importance entre le contexte rural et urbain. Ainsi, la proportion de logement sans voiture, ou celle de maisons individuelles, n'a pas la même interprétation dans un contexte d'habitat urbain dense ayant des transports en commun ou dans un contexte rural où la voiture est presque indispensable. Une solution qui a été adoptée est alors de distinguer les IRIS suivant différents classes d'urbanisation et d'appliquer la procédure séparément sur chacune de ces classes.

C'est également ce type d'approche qui a été testé dans un autre projet s'intéressant aux liens entre les caractéristiques socio-économiques et l'accès au soin chez les femmes enceintes dans les départements de Seine-et-Marne (77) et de Seine-Saint-Denis (93), ce qui nous a permis d'évaluer l'intérêt de cette procédure dans des contextes différents.

7.7 Discussion

Dans ce chapitre, nous avons présenté une procédure de création d'indices socio-économiques que nous avons développée durant cette thèse. Cette procédure se base sur les techniques d'analyses de données, en particulier l'ACP, afin d'avoir une optique de sélection et de synthèse « guidée par les données ». Nous avons également étudié différentes méthodes pour discrétiser l'indice quantitatif construit et développé un package R implémentant cette procédure. Nous avons illustré cette procédure sur les différentes agglomérations du projet Equit'Area et présenté certains de ces résultats.

Comme nous l'avons détaillé en section 7.2, cette procédure a été conçue pour être aussi flexible et reproductible que possible à différents contextes d'application. Plutôt que de proposer une formule définie d'une sélection de variables choisies arbitrairement, elle effectue à partir d'un vaste ensemble de variables (sélectionnées à partir de la littérature) une première sélection sur des critères statistiques, puis synthétise l'ensemble à l'aide d'une technique adaptée.

Le nombre de variables sélectionnées, non défini *a priori*, est stable entre les différentes analyses et est d'environ 20. Cette diversité permet à la fois de déterminer les variables communes entre agglomérations et spécifiques à celles-ci, mais aussi d'avoir une interprétation de l'indice et de ses valeurs dans les différents IRIS.

Comme tous les indices, celui créé par notre procédure a comme limite d'être un indicateur composite dont l'interprétation peut être compliquée par rapport à des variables socio-économiques couramment utilisées. Par ailleurs, comme nous l'avons évoqué en section 7.3, il s'agit d'un indice relatif qui ne permet de comparer que des IRIS (ou autres unités géographiques) faisant partie de la zone sur laquelle la procédure a été appliquée. Ainsi, les comparaisons (tests d'égalité des moyennes ou de Wilcoxon sur échantillons appariés) ayant été réalisées au cours de ce travail entre les indices spécifiques à chaque

agglomération et l'indice « global » montrent des différences significatives. La comparaison de l'indice en tant que tel entre différents pays ou différentes années peut alors être compliquée. Néanmoins, l'application de l'ensemble de la procédure, incluant notamment les résultats sur les variables sélectionnées, peut permettre des comparaisons sur la composition de l'indice à défaut de permettre de comparer les zones elles-mêmes.

La procédure présentée ici est par ailleurs plus complexe à mettre en œuvre que la création d'autres indices. Par rapport à ces indices (et notamment ceux de Townsend et Carstairs), cette procédure :

- permet d'avoir, grâce à sa sélection de variables sur des critères statistiques, une approche avec aussi peu d'a priori que possible plutôt que de devoir effectuer une sélection arbitraire -généralement basée sur des choix parmi les variables reconnues dans la littérature-.
- n'est pas restreinte dans son utilisation si l'on ne dispose pas de certaines variables, grâce au fait qu'il s'agit d'une méthode de construction d'indices plutôt que d'une formule « toute faite ». Ainsi, ne pas disposer d'une variable indiquant le surpeuplement des logements empêche de calculer les indices de Townsend et Carstairs, mais pas d'appliquer notre procédure.
- autorise une interprétation plus fine des résultats grâce au nombre important de variables sélectionnées dans les indices. Il est ainsi possible, après avoir identifié les zones défavorisées à l'aide de l'indice, d'observer plus précisément les caractéristiques et particularités de ces zones vis-vis des variables composant l'indice.

Ces différentes raisons sont ainsi autant d'apports pouvant justifier l'utilisation d'une procédure plus complexe malgré une corrélation importante avec les indices existants

La classification par CAH étudiée, bien que classique d'un point de vue statistique, n'a été que rarement utilisée dans ce domaine à notre connaissance. Elle permet, à l'inverse des quantiles utilisés généralement, de créer des classes d'IRIS plus homogènes. Elle a cependant comme inconvénient de n'avoir aucun critère de taille des classes, ce qui peut poser certains problèmes dans les applications.

La CAH paraît donc intéressante d'un point de vue descriptif pour créer des classes de composition homogène, par exemple afin d'identifier des « points noirs » et d'aider les décideurs locaux dans leurs politiques sociales ou l'aménagement du territoire. Cependant, notre réflexion sur ses apports dans le cadre d'études explicatives est encore en cours, sa plus grande complexité de mise en œuvre et d'interprétation et surtout la possibilité d'obtenir des classes d'effectifs très différents pouvant être une limite dans des modèles explicatifs.

La diversité des applications présentées dans la section précédente nous a permis d'avoir de multiples informations sur les sélections de variables réalisées et les particularités des différentes zones, mais aussi sur la reproductibilité et la stabilité de la procédure. Il s'avère, jusqu'à présent, que les variables sélectionnées comportent généralement un important « tronc commun » (similaire à celui visible dans le tableau 7.2) et que la procédure a systématiquement permis d'opposer le long du premier axe de l'ACP finale des variables de « faveur » à des variables de « défaveur ».

Chapitre 8

Indice de multi-expositions environnementales

8.1 Contexte

À chaque instant, toute personne est exposée à un mélange de nombreuses pollutions, nuisances ou aménités environnementales dans son cadre de vie. Ces expositions peuvent être de nature et de source très diverses. Certaines sont reconnues pour avoir des effets négatifs sur la santé (maladies cardiovasculaires et respiratoires, cancers, issues de grossesse défavorables, problèmes mentaux, ...) comme par exemple la pollution de l'air (issue du trafic, du chauffage urbain ou de l'industrie)[33, 109], la contamination des eaux et des sols [122], le bruit (de trafic routier, aérien, ferroviaire, d'industries)[119, 145] ou encore la proximité de décharges ou d'industries dangereuses [122] ... D'autres expositions ont un effet bénéfique sur la santé, comme par exemple la proximité aux espaces verts [127, 128].

Malgré cela, la plupart des études épidémiologiques s'intéressant aux liens entre santé et environnement ne prennent en compte qu'une seule exposition afin d'établir son impact sur la santé. Un appel pour des approches plus « réalistes » existe cependant depuis quelques années pour que les cumuls d'expositions multiples soient pris en compte. Cet appel provient autant des scientifiques [146–148] que des différents rapports et directives publics [149–152]. L'une des possibilités pour répondre à cet appel est d'essayer d'estimer le « fardeau environnemental » qui pèse sur les populations dans leur cadre de vie en utilisant un indice de multi-exposition.

Considérer plusieurs expositions simultanément pose cependant différents problèmes méthodologiques. Ainsi, les différentes expositions sont souvent corrélées car elles partagent des sources communes (le trafic routier par exemple qui induit bruit et pollution de l'air) ce qui peut poser des problèmes si elles sont directement intégrées dans des modèles et justifie de chercher à les synthétiser par un indice composite.

Mais les variables représentant les expositions sont généralement de types variés, pouvant être aussi bien qualitatives que quantitatives, avec différentes unités de mesures ($\mu g/m^3$, dB, %, ...). Par ailleurs, le choix des poids à affecter à chaque exposition dans la construction d'un tel indice est souvent difficile (en particulier si l'on souhaite prendre en compte l'effet sur la santé de chacune).

Bien que résoudre certaines de ces difficultés puisse sembler simple du point de vue d'un statisticien, le domaine de l'épidémiologie environnementale pointe le manque d'outils pour évaluer la multi-expositions, ce qui peut également illustrer l'importance de transférer les développements de mathématiques appliquées aux applications. En effet, les études définissant de tels indices d'expositions multiples sont peu nombreuses et ne prennent souvent en compte qu'une seule famille de polluants ayant les mêmes effets sanitaires (notamment la pollution de l'air car ses effets et ses modes d'actions sont connus depuis longtemps) qui est agrégée ensuite à l'aide de sommes ou de produits pondérés, ou encore de scores définis arbitrairement [147, 153–156].

Par exemple, Pearce *et al.* [147] définissent un indice de multi-expositions environnementales en Nouvelle-Zélande à une échelle similaire à celle de l'IRIS (les CAU, *Census Area Units*). Ils sélectionnent d'abord des expositions estimées négatives (pollution de l'air, climat froid) ou positives (exposition aux UV, espaces verts) pour la santé puis ils attribuent un score de :

- +1 pour les quartiers parmi les 20% les plus exposés à la pollution de l'air,
- +1 pour les quartiers parmi les 20% les plus exposés au climat froid,
- -1 pour les quartiers parmi les 20% les plus exposés aux UV,
- -1 pour les quartiers parmi les 20% les plus exposés aux espaces verts.

L'indice de multi-exposition est alors la somme de ces quatre scores et varie donc entre -2 à +2.

Dans ce contexte et dans la mesure où de nombreuses variables environnementales peuvent être disponibles, l'utilisation de techniques d'analyse de données semble peu employée bien que parfaitement adaptée. C'est donc ces techniques que nous avons employées pour créer un indice de multi-expositions environnementales. Comme nous avons pu le détailler précédemment (voir chapitre 6), les données dont nous disposons ici sont de qualités et de précisions variées. Par ailleurs, nous sommes ici dans un cas où il apparaît impossible d'établir une hiérarchie entre les différents profils d'expositions : par exemple, il n'est en rien évident de dire si un IRIS très exposé à la pollution de l'air mais pas aux industries polluantes est « pire » qu'un IRIS peu exposé au NO_2 mais très proche de nombreuses industries polluantes. Pour ces deux raisons, nous avons donc choisi de construire un indice de multi-expositions qualitatif nominal

8.2 Construction de l'indice de multi-expositions

Face à l'importante base de données d'expositions environnementales dont nous disposons, nous nous sommes rapidement dirigés vers les techniques d'analyse de données, qui semblaient adaptées au problème. Il nous a alors fallu identifier quelle était la meilleure méthode à utiliser étant donné la nature des variables environnementales et l'objectif visé.

> Ici, contrairement au chapitre précédent, nous souhaitons étudier le cumul d'expositions de différents types. En effet, les différentes expositions environnementales sont naturellement réparties en plusieurs groupes (ceux-ci ont été présentés dans le chapitre 6 : pollution de l'air, bruit, proximité aux industries polluantes, proximité au trafic et aux espaces verts). Puisque l'on souhaite ici que chaque groupe d'exposition soit représenté dans l'indice final, cela nous a conduits à choisir de donner à chacun un poids identique (tout en gardant la diversité d'indicateurs présente dans chaque groupe).

> Ceci, associé à la nature mixte des groupes (composés soit de variables quantitatives, soit de variables qualitatives), nous a naturellement menés vers l'analyse factorielle multiple (AFM). Cette méthode, définie par Escofier et Pagès [157], permet en effet d'étudier simultanément des groupes de variables quantitatives et qualitatives, tout en donnant à chaque groupe un poids identique dans l'analyse (indépendamment du nombre de variables de chacun).

À notre connaissance, l'AFM n'a jamais été utilisée dans un tel contexte. Ceci peut être lié à plusieurs raisons. D'une part, de la même manière que nous l'avons évoqué en section 7.1 pour l'ACP, la succession de difficultés de définition et de choix des expositions, de modélisation et/ou récolte des données, de conception de l'étude, ... peut conduire à utiliser des techniques connues et « traditionnelles » dans le domaine plutôt que de s'engager dans une nouvelle réflexion méthodologique sur la définition d'un nouvel indicateur de multi-expositions.

D'autre part, dans l'étude d'expositions multiples il est souvent recherché d'inclure en partie l'effet sur la santé de chaque exposition. Il faut ainsi rechercher ou établir des relations dose-réponse capables de donner des indications précises sur l'effet sanitaire de toutes les expositions, puis trouver comment combiner ces informations en un indicateur de multi-expositions. C'est notamment l'une des raisons expliquant que la plupart des indices de multi-expositions existant s'intéressent à des expositions de natures identiques et dont les effets physiologiques sont similaires, et l'utilisation de l'AFM ne semble pas être pertinente dans ce cadre précis.

Cependant, nous poursuivons ici un objectif différent puisque nous souhaitons établir un indicateur de « pression environnementale » incluant des variables de natures variées et d'effets divers sur la santé, sans inclure forcément l'effet précis de chaque exposition. L'AFM apparaît alors comme une solution appropriée pour ce travail.

Une sélection des variables à inclure dans l'AFM a d'abord été réalisée. Le but est de choisir les variables de chaque groupe qui semblent comme étant les plus intéressantes pour décrire les contrastes environnementaux dans les zones d'études. En effet, la collecte de ces données ayant un coût important (temporel, humain et financier), on souhaite savoir s'il est possible de synthétiser un type d'exposition avec moins de variables pour une meilleure reproductibilité de l'indice.

En se basant sur le même principe que la première étape de la procédure de création d'indices socio-économiques détaillée en section 7.2.1, une ACP ou une analyse factorielle des correspondances multiples (AFCM), selon la nature des variables, est réalisée sur chaque groupe. Contrairement à la première étape de la procédure de création d'indices socio-économiques, le nombre de variables conservées ici dans chaque groupe n'est pas fixé *a priori* mais est fonction des résultats de l'analyse et de l'intérêt que peut avoir chaque variable.

En effet, alors que dans le cas de la procédure de création d'indices socio-économiques les groupes de variables « redondantes » étaient formés de variables représentant la même notion et fortement corrélées entre elles, les groupes d'expositions environnementales représentent parfois différents indicateurs ou différentes manières d'évaluer une exposition. Ces indicateurs ont par conséquent des corrélations généralement moins importantes, et

peuvent apporter des informations différentes. Pour cette raison, nous avons effectué un choix raisonné à partir des résultats des analyses et des particularités de chaque groupe d'expositions. Ainsi, dans le cas du NO_2 les différentes variables représentent la même mesure faite à différentes années et sont très corrélées, ce qui conduit à n'avoir qu'un facteur de taille lorsque l'on effectue une ACP sur ces variables et donc à ne conserver que des variables corrélées avec le premier facteur. A contrario, les indicateurs de bruit sont construits suivant différentes méthodes, ce qui conduit à l'apparition de deux axes intéressants avec une ACP et donc à garder des variables corrélées avec les deux premiers facteurs.

Le but de cette première sélection est également de faciliter l'interprétation de l'indice final en éliminant les variables trop « redondantes » tout en gardant au moins deux variables par groupe d'exposition.

Une fois cette sélection réalisée, on applique l'AFM à cette sélection puis une CAH sur les facteurs obtenus ce qui nous permet d'obtenir un indice qualitatif basé sur des critères statistiques et faisant la synthèse des informations apportées par les différentes expositions. Là encore, la CAH n'a, à notre connaissance, jamais été utilisée dans ce contexte, probablement pour des raisons similaires à celles détaillées plus haut dans le cas de l'AFM.

8.3 Résumé des principaux résultats

L'ensemble des données environnementales considérées n'étant disponible que pour l'agglomération de Lyon actuellement, l'indice de multi-expositions n'a été construit que sur celle-ci. La table 8.1 résume les variables environnementales disponibles et celles qui ont été retenues pour être intégrées à l'AFM.

On a ensuite appliqué l'AFM sur les variables choisies réparties suivant les groupes indiqués précédemment. Les quatre premiers facteurs expliquent respectivement environ 30%, 16%, 12% et 9% de la variance totale.

Étant donné les contributions des groupes et des variables (ainsi que les corrélations pour les variables quantitatives) (voir tables 8.2 et 8.3) il est possible d'interpréter ces quatre premiers facteurs de la façon suivante :

- Le premier facteur a pour plus importants contributeurs les groupes de pollution de l'air et d'exposition au trafic automobile. Il est fortement corrélé positivement ($\geq$0,80) aux deux indicateurs de NO_2 et de trafic, et corrélé négativement aux deux indicateurs d'espaces verts, en particulier la surface d'espaces verts. Cela semble donc être un axe qui va opposer les IRIS de petite taille, peu verts avec une population proche des voies à fort trafic et exposée à la pollution de l'air, aux IRIS plus grands, plus verts et moins exposés au NO_2.

- Le second facteur est justifié par les industries. Les modalités 0 des indicateurs de proximité aux industries ont une coordonnée négative sur cet axe tandis que les modalités 1 ont des coordonnées largement positives. Il semble donc que cet axe reflète principalement l'opposition entre les IRIS ayant des industries polluantes dans leur proximité et ceux n'en ayant pas.

- Les groupes contribuant le plus au troisième facteur sont les niveaux de bruit et les espaces verts. L'ensemble des indicateurs de bruit présente une corrélation positive moyenne ($\sim$ 0,55) avec ce facteur, et c'est également le cas de la proportion d'espaces verts. Les IRIS à la fois exposés à des niveaux plus importants de bruit et ayant une

TABLE 8.1 – Variables environnementales disponibles et sélectionnées, par type d'exposition.

Type d'exposition	Description
Pollution de l'air (NO_2)	Concentration moyenne annuelle 2002 ($\mu g/m^3$)
	Concentration moyenne annuelle 2003 ($\mu g/m^3$)
	Concentration moyenne annuelle 2004 ($\mu g/m^3$)
	Concentration moyenne annuelle 2005 ($\mu g/m^3$)
	Concentration moyenne annuelle 2006 ($\mu g/m^3$)
	Concentration moyenne annuelle 2007 ($\mu g/m^3$)
	Concentration moyenne annuelle 2008 ($\mu g/m^3$)
	Concentration moyenne annuelle 2009 ($\mu g/m^3$)
Bruit (L_{DEN})	Moyenne énergétique (dB(A))
	Moyenne arithmétique (dB(A))
	Médiane (dB(A))
	Moyenne énergétique prenant en compte la population des bâtiments (dB(A))
	Moyenne arithmétique prenant en compte la population des bâtiments (dB(A))
	Médiane prenant en compte la population des bâtiments (dB(A))
Proximité aux industries	Présence d'au moins une industrie polluante (Modalités : 0, 1)
	Nombre d'industries polluantes (Modalités : 0, 1, 2)
	Présence d'au moins un buffer de 500m de rayon (Modalités : 0, 1)
	Nombre de buffers de 500m de rayon (Modalités : 0, 1, 2+)
	Présence d'au moins un buffer de 1km de rayon (Modalités : 0, 1)
	Nombre de buffers de 1km de rayon (Modalités : 0, 1, 2, 3+)
Proximité au trafic	Proportion de la population à moins de 100m d'une voie à fort trafic (%)
	Proportion de la population à moins de 150m d'une voie à fort trafic (%)
	Proportion de la population à moins de 200m d'une voie à fort trafic (%)
	Proportion de la population à moins de 250m d'une voie à fort trafic (%)
	Proportion de la population à moins de 300m d'une voie à fort trafic (%)
Espaces verts	*Surface verte totale (m^2)*
	Proportion d'espaces verts parmi la surface totale (%)

Toutes les variables sont à l'échelle de l'IRIS

Italique : variables sélectionnées pour composer l'indice multi-expositions

proportion d'espaces verts importante sont ainsi opposés sur cet axe aux IRIS moins bruyants et proportionnellement moins verts.
L'examen des IRIS semble indiquer une opposition entre les grands (et plus verts) IRIS de périphérie, proches d'autoroutes (et donc exposés au bruit) et les IRIS plus petits et/ou plus urbains, parfois avec peu de population (donc des niveaux de bruit pondérés par la population plus faibles).

– Le quatrième facteur a pour contributeur majoritaire le groupe d'exposition au bruit. Plus particulièrement, il est fortement corrélé négativement ($\sim -0{,}70$) aux variables de bruit prenant en compte la population des bâtiments. Cet axe oppose donc essentiellement les IRIS peu peuplés ayant des valeurs faibles ou nulles pour les variables précédentes, aux IRIS dont la population est plus exposée au bruit.

TABLE 8.2 – Contributions des groupes aux quatre premiers axes de l'AFM

Groupes	**Axe 1**	**Axe 2**	**Axe 3**	**Axe 4**
Pollution de l'air	30,00	4,67	1,03	19,34
Bruit	13,30	13,64	57,08	70,05
Proximité aux industries	0,11	81,25	6,55	4,02
Proximité au trafic	33,57	0,03	0,92	0,22
Espaces verts	23,02	0,41	34,42	6,36

Contribution des groupes aux axes (en %)

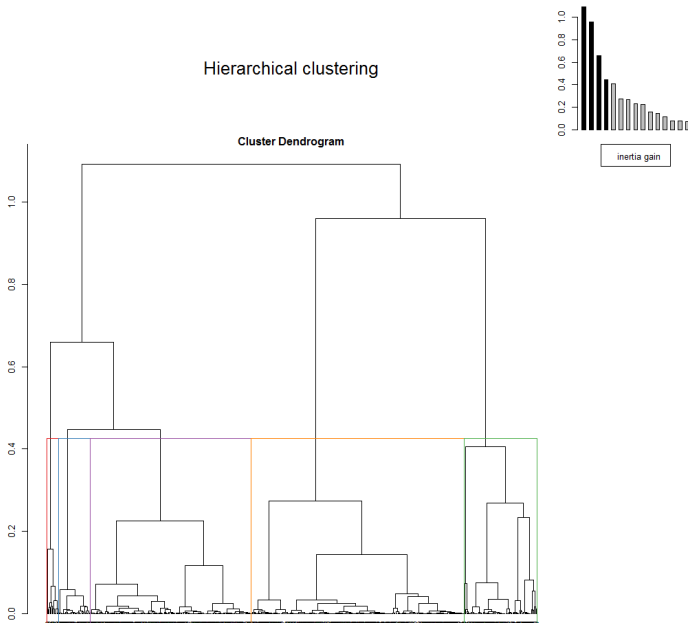


FIGURE 8.1 – Dendrogramme de la CAH sur les facteurs de l'AFM.

TABLE 8.3 – Corrélations et contributions des variables aux quatre premiers facteurs de l'AFM

Variables		Axe 1		Axe 2		Axe 3		Axe 4	
		Coord[a]	Ctr[b]	Coord[a]	Ctr[b]	Coord[a]	Ctr[b]	Coord[a]	Ctr[b]
NO_2_2003		0,8	15,05	−0,23	2,32	0,09	0,46	0,36	9,61
NO_2_2006		0,8	14,95	−0,23	2,36	0,1	0,57	0,37	9,74
Bruit_moyenne_arith		0,36	2,51	0,44	7,13	0,56	15,15	0,34	6,91
Bruit_médiane		0,44	3,69	0,36	4,99	0,55	14,31	0,35	7,3
Bruit_moyenne_energ_pop		0,42	3,47	0,12	0,5	0,53	13,57	−0,69	28,58
Bruit_moyenne_arith_pop		0,43	3,63	0,17	1,02	0,54	14,04	−0,68	27,26
Trafic_200m		0,84	17,02	−0,02	0,03	−0,09	0,45	0,00	0,00
Trafic_250m		0,83	16,55	0,00	0,00	−0,09	0,47	−0,06	0,22
Espaces vert_surface		−0,68	13,77	0,05	0,15	0,4	11,5	0,16	2,44
Espaces vert_proportion		−0,56	9,26	−0,07	0,26	0,57	22,92	0,21	3,92
Industries_nombre	0	0,01	0,00	−0,12	0,39	0,02	0,02	−0,01	0,01
	1	−0,57	0,07	2,85	6,12	−0,44	0,23	−0,25	0,12
	2	0,42	0,02	3,31	3,82	−0,54	0,16	1,63	2,35
Industries_500m_présence	0	−0,01	0,00	−0,4	3,97	0,09	0,31	−0,02	0,03
	1	0,07	0,01	2,2	21,73	−0,49	1,69	0,13	0,18
Industries_500m_nombre	0	−0,01	0,00	−0,4	3,97	0,09	0,31	−0,02	0,03
	1	0,13	0,01	2,05	13,66	−0,61	1,96	−0,05	0,02
	2+	−0,07	0,00	2,62	8,36	−0,15	0,04	0,61	1,14
Industries_1km_présence	0	−0,01	0,00	−0,52	5,51	0,13	0,53	0,03	0,04
	1	0,02	0,00	1,29	13,72	−0,31	1,31	−0,06	0,09

[a]Coordonnée de la variable sur l'axe, i.e. coefficient de corrélation entre la variable et le facteur (pour les variables quantitatives)

[b]Contribution de la variable ou de la modalité à l'axe (en %)

On choisit de couper l'arbre de la CAH (figure 8.1) de façon à avoir une partition en cinq classes. Ce nombre a été choisi après l'examen du dendrogramme mais également à partir de l'interprétation des différentes classes. Ainsi, si une partition en trois classes apparaissait comme préférable du point de vue de l'inertie, ce nombre semblait trop petit et la classification des IRIS obtenue trop grossière pour présenter un intérêt dans les applications (voir figure 8.2a). Il en était de même pour la partition en quatre classes, c'est pourquoi la partition en cinq classes a finalement été retenue. Celles-ci peuvent être interprétées de la manière suivante (voir tables 8.4 et 8.5 et figure 8.2b) :

- Une première classe composée uniquement d'IRIS de type « divers » (voir section 3.1.2) sans population. On y trouve des zones d'activité ou industrielles à la périphérie de Lyon, ou encore la Préfecture, un Hôpital ou la Cité Internationale à l'intérieur de Lyon. Les IRIS de cette classe ont une très grande surface verte en moyenne, et des indicateurs prenant en compte la population (bruit, trafic) très faibles voire nuls, ce qui est cohérent avec la définition des IRIS divers (de grandes zones peu peuplées).
- Une seconde classe regroupe des IRIS composés de zones pavillonnaires, de maisons avec jardin, de maisons isolées. Peu exposée au trafic, à la pollution de l'air et au bruit, avec une très forte proportion d'espaces verts, il s'agit de la classe la plus « rurale ».

- La troisième classe est composée exclusivement d'IRIS à moins de 500m d'une industrie polluante, souvent composés soit de grands ensembles d'immeubles, soit de zones avec à la fois de nombreux hangars ou industries mêlés à des maisons.
- La quatrième classe regroupe des IRIS de zones pavillonnaires plus proches des centres-villes. Tout comme la classe 2 ils sont moins exposés au bruit, à la pollution de l'air ou au trafic que la moyenne des autres IRIS mais dans une moindre mesure que la classe 2.
- La cinquième classe comprend quasiment toute la commune de Lyon et ses alentours immédiats, ainsi que quelques centres-villes d'autres communes. Les IRIS qui la composent sont beaucoup plus exposés au bruit, à la pollution de l'air et au trafic, et ont une faible proportion d'espaces verts. Il s'agit d'une classe très « urbaine ».

Ces classes permettent de déterminer différents profils d'exposition existant dans le Grand Lyon, de les expliciter et de les localiser.

TABLE 8.4 – Valeurs moyennes des variables quantitatives, par classe

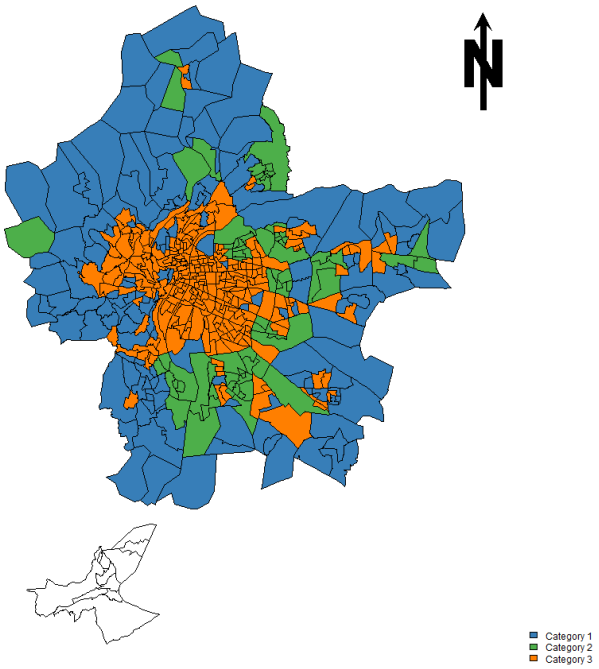
Variables	**Classe 1** *(n=13)*	**Classe 2** *(n=35)*	**Classe 3** *(n=70)*	**Classe 4** *(n=151)*	**Classe 5** *(n=230)*	**Total** *(n=499)*
NO_2_2003[a]	39,94	32,47	39,64	36	44,56	40,31
NO_2_2006[a]	42,04	35,11	41,77	38,38	46,45	42,44
Bruit_moyenne_arith[b]	63,03	62,67	65,69	62,87	64,29	63,91
Bruit_médiane[b]	62,97	62,47	65,77	62,69	65,1	64,22
Bruit_moyenne_energ_pop[b]	0,00	67,79	69,51	66,21	70,7	67,13
Bruit_moyenne_arith_pop[b]	0,00	63,91	67,41	63,56	67,34	64,21
Espaces_vert_surface[c]	610 963	1 447 667	128 167	126 383	32 232	188 536
Espaces_vert_proportion[d]	16,45	34,32	8,53	13,06	8,48	11,89
Trafic_200m[d]	52,77	35,72	82,98	59,2	95,6	77,5
Trafic_250m[d]	53,78	43,28	89,25	69,71	98,48	83,44

n : nombre d'IRIS dans la classe
[a] $\mu g/m^3$; [b] dB(A) ; [c] m^2 ; [d] %

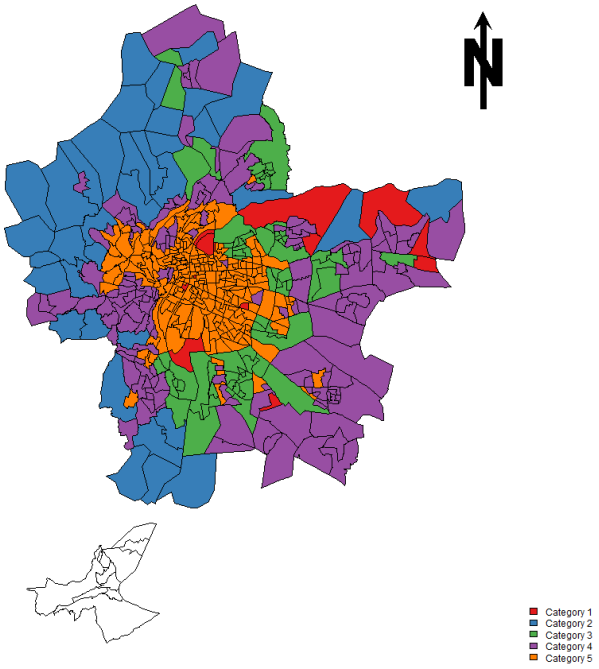
TABLE 8.5 – Répartition des modalités des variables qualitatives par classes

Variables	**Modalités**	**Classe 1** *(n=13)*	**Classe 2** *(n=35)*	**Classe 3** *(n=70)*	**Classe 4** *(n=151)*	**Classe 5** *(n=230)*	**Total** *(n=499)*
Industries_nombre	0	12 (92%)	34 (97%)	53 (76%)	151 (100%)	230 (100%)	480 (96%)
	1	0 (0%)	1 (3%)	12 (17%)	0 (0%)	0 (0%)	13 (3%)
	2	1 (8%)	0 (0%)	5 (7%)	0 (0%)	0 (0%)	6 (1%)
Industries_500m_présence	0	10 (77%)	32 (91%)	0 (0%)	150 (99%)	230 (100%)	422 (85%)
	1	3 (23%)	3 (9%)	70 (100%)	1 (1%)	0 (0%)	77 (15%)
Industries_500m_nombre	0	10 (77%)	32 (91%)	0 (0%)	150 (99%)	230 (100%)	422 (85%)
	1	2 (15%)	2 (6%)	52 (74%)	0 (0%)	0 (0%)	56 (11%)
	2+	1 (8%)	1 (3%)	18 (26%)	1 (1%)	0 (0%)	21 (4%)
Industries_1km_présence	0	10 (77%)	28 (80%)	70 (100%)	122 (81%)	196 (85%)	426 (85%)
	1	3 (23%)	7 (20%)	0 (0%)	29 (19%)	34 (15%)	73 (15%)

n (x%) : n est le nombre d'IRIS avec la modalité. x est le pourcentage d'IRIS avec la modalité dans la classe



(a) Partition en 3 classes



(b) Partition en 5 classes

FIGURE 8.2 – Carte des catégories de multi-expositions pour le Grand Lyon.

8.4 Article : Data Analysis Technics, a Tool for Cumulative Exposure Assessment

Un article présentant la construction de l'indice de multi-expositions exposée dans les sections précédentes et son application comme un exemple de l'utilité des techniques d'analyse de données dans le champ de l'évaluation d'expositions multiples est en cours d'écriture. Il sera soumis prochainement au *Journal of Exposure Science and Environmental Epidemiology*. Son résumé est présenté en annexe B.3.

8.5 Discussion

Dans ce chapitre, nous avons exposé la méthode que nous avons choisi d'utiliser pour répondre à l'objectif de mise au point d'indicateurs composites synthétisant le « fardeau environnemental » pesant sur les populations.

L'indice de multi-expositions qualitatif que nous créons permet ainsi de déterminer différents profils d'exposition. À notre connaissance, ce travail est l'un des rares dans le domaine de l'épidémiologie environnementale à utiliser des techniques d'analyses de données, en particulier sur des expositions aussi diversifiées et à une échelle géographique aussi fine. L'utilisation de l'analyse de données permet ainsi d'avoir une approche « guidée par les données » avec peu d'*a priori* sur le choix des variables à intégrer dans l'indice et autorise, comme dans le cas de l'indice socio-économique présenté au chapitre précédent, à revenir aux variables pour interpréter les résultats et les classes obtenues plus clairement.

L'indice qualitatif obtenu permet, en étant utilisé seul, de discerner rapidement les différents profils d'expositions environnementales, ce qui peut aider les décideurs public à situer les lieux ayant un fardeau environnemental plus important et à adapter leur planification urbaine en conséquence. Dans le cas où les données environnementales seraient disponibles périodiquement, notre approche peut également permettre d'estimer l'évolution temporelle soit des IRIS entre les différentes catégories, soit des catégories elles-mêmes. Cet indice peut par ailleurs être intégré directement dans des modèles explicatifs, que ce soit dans des domaines comme l'épidémiologie environnementale ou la justice environnementale.

Bien qu'il semble en lui-même intéressant dans un contexte descriptif, cet indice de multi-expositions a différentes limites. En tant qu'indice, il possède les limites communes aux indices et déjà exposées en section 7.7 : la complexité potentielle d'interprétation ou encore la relativité de la construction des classes.

De plus, il semble à l'heure actuelle que cet indice ne permette pas de mettre en évidence des effets sur la mortalité infantile dont on sait par ailleurs qu'ils existent lorsque l'on utilise les expositions individuellement. Ainsi, le travail présenté en annexe C.3 a permis de mettre en évidence à Lille et Lyon, à l'aide de modèles additifs généralisés, une influence du dioxyde d'azote sur la mortalité infantile. De même, des liens entre le bruit et la mortalité infantile (annexe C.5) et entre les espaces verts et la mortalité néonatale (annexe C.6) ont été mis en évidence, toujours à Lyon, à l'aide d'analyses spatiales (de type balayage spatial).

Comme nous le détaillerons davantage en section 10.3, nous n'avons pas pu mettre en évidence un tel lien avec la mortalité infantile en utilisant l'indice de multi-expositions présenté ici et le modèle « BYM ». Ceci pourrait être dû à différentes raisons :

- Tout d'abord, il pourrait être trop synthétique et/ou trop grossier, et par conséquent « gommer » les effets des différentes expositions individuelles. Dans ce cas, les variables et le nombre de classes retenues pourraient être changés, et la méthode de construction de l'indice elle-même pourrait être revue, afin d'obtenir un indice plus fin.

- Une autre raison, opposée à la première, pourrait être que cet indice ne présente pas encore suffisamment de variété d'expositions pour donner une bonne estimation du cumul d'expositions et de son effet sur la santé. Dans ce cas, l'ajout d'autres types d'expositions (pollution des sols, par exemple), pourrait permettre de compléter la mesure du « fardeau environnemental » et mettre en évidence son influence sur la santé.

- Enfin, il est également possible que sur la zone ayant servi d'illustration, il n'existe pas d'effet du cumul d'expositions sur la santé, ou que cet effet soit modeste, et s'efface dès lors qu'est également pris en compte le profil socio-économique des IRIS. Dans ce cas, d'autres tests sur d'autres agglomérations pourraient révéler un tel effet, sans changements dans la méthode ou les données employées.

L'ensemble de ces différentes possibilités fait toujours partie des réflexions que nous menons encore sur le sujet. Ainsi, la collecte de davantage de données environnementales (les anciens sites et sols pollués, par exemple), à la fois dans l'agglomération de Lyon mais sur les autres agglomérations du projet également, devrait permettre de tester cet indice sur d'autres zones et avec des expositions environnementales plus diversifiées.

Parmi les autres limites de cet indice, nous pouvons également citer le parti pris de donner un poids identique à chaque type d'exposition. Si cette décision est apparue naturellement dans notre contexte d'application et étant donné les informations à notre disposition, il existe d'autres contextes où l'on souhaite donner aux différentes expositions un poids proportionnel à leur impact connu sur la santé afin d'avoir une évaluation « plus réaliste » du risque sanitaire encouru par les populations. Ce point est cependant particulièrement complexe à résoudre, puisqu'il implique d'une part de connaître ou d'évaluer l'impact de chaque exposition sur la santé (ce qui peut être très compliqué), puis de déterminer les poids relatifs de chaque type d'exposition en fonction de cet impact. Ceci dépasse largement le cadre de la présente thèse et nécessiterait une étroite collaboration entre experts en mesure des expositions, cliniciens, biologistes, épidémiologistes et statisticiens.

Chapitre 9

Analyse de données multi-niveaux

9.1 Contexte

Ainsi que cela a été détaillé dans les parties précédentes, le choix a été fait de travailler à l'échelle des IRIS, qui représentent donc soit des quartiers soit des petites communes, afin d'utiliser la plus fine échelle géographique disponible.

Néanmoins, il est souvent possible d'avoir simultanément des données à plusieurs échelles géographiques. On peut avoir à la fois des informations directement sur l'individu et sur son IRIS de résidence, ou encore avoir des informations sur l'IRIS et sur la commune à laquelle il appartient. Or, chacune de ces échelles peut avoir une influence sur la santé. C'est le cas pour le statut socio-économique pour lequel des études ont montré que s'il avait une influence sur de nombreux événements de santé à l'échelle de l'individu, il demeurait également une influence du statut socio-économique contextuel (du quartier, de la commune, ...) sur celle-ci même après prise en compte de l'effet individuel [3, 10, 158].

La prise en compte de ce genre de données multi-niveaux nécessite l'utilisation de techniques statistiques spécifiques, mais elle est de plus en plus fréquente en épidémiologie. Souvent, celle-ci se fait sur la base d'une analyse de régression (linéaire, logistique, de Poisson, de Cox, ...) adaptée à ce type de données dans laquelle les effets des différents niveaux sont modélisés. Cependant, dans la même optique que celle exposée dans les chapitres précédents, effectuer une exploration guidée par les données à l'aide de techniques d'analyse de données multi-niveaux semble moins fréquent.

Dans la mesure où il est également possible d'obtenir des données à différentes échelles dans notre cadre d'application, nous avons donc tenté d'employer ce type d'analyse à l'étude de données socio-économiques simultanément à l'échelle de la commune et de l'IRIS. L'objectif est alors de **déterminer s'il est possible de constituer des indicateurs composites prenant en compte les différents niveaux géographiques des données**, ainsi que de déterminer au sein des agglomérations quels seraient les déterminants socio-économiques existant à l'échelle des communes et ceux existant à l'échelle des IRIS une fois l'effet des communes retiré.

9.2 Présentation de la méthode

Dans le contexte expliqué précédemment, nous avons donc employé une méthode d'analyse de données présentée par Lebart, Piron et Morineau [159]. Cette technique nous a paru particulièrement justifiée ici puisqu'elle permet d'appliquer sur des données structurées (en particulier avec une structure de partition) les techniques classiques d'analyse de données. Ceci nous permet d'appliquer la même procédure que celle présentée au chapitre 7.2 mais en tenant également compte du niveau communal.

Considérons X une table de données de n lignes représentant les individus statistiques et p colonnes représentant les variables (on considère que celles-ci sont centrées). On va considérer de plus que l'ensemble des individus est partitionné en q classes connues. Dans notre cas, les n individus sont ainsi les IRIS, répartis en q communes. On va effectuer successivement une analyse inter-classes qui va étudier les positions relatives des classes sans tenir compte des individus puis une analyse intra-classes qui décrira les positions relatives des individus au sein de leur classe.

Analyse inter-classes

L'analyse inter-classes consistera à appliquer une technique d'analyse de données au tableau agrégé de q lignes et p colonnes.

Dans notre cas, il s'agira donc d'appliquer la procédure présentée dans la section 7.2.1 sur les mêmes variables que celles présentées dans le chapitre 5 mais construites à l'échelle des communes.

Analyse intra-classes

L'analyse intra-classes sera l'application de la même technique d'analyse de données sur un tableau de n lignes et p colonnes mais où les données auront été centrées par commune.

Si x_i^j représente la valeur de la variable j pour l'individu i et que ce dernier appartient à la classe k, on considérera dans l'analyse intra-classes la valeur $x_i^j - \bar{x}_k^j$ où $\bar{x}_k^j$ représente la valeur de la variable j pour la classe k (dont on dispose soit directement, soit en utilisant la moyenne des valeurs des individus appartenant à la classe k).

Là encore, dans notre cas, il s'agira d'appliquer la procédure de création d'indices socio-économiques présentée précédemment mais sur les données des IRIS centrées par leur commune d'appartenance.

9.3 Résumé des principaux résultats

Un test d'application de cette méthode a été réalisé pour l'agglomération du Grand Lyon avec les données du recensement de 2006. Les données utilisées sont les mêmes variables socio-économiques que celles employées dans la section 7.3, exception faite de quelques variables qui n'ont pas pu être construites à l'échelle communale avec les données disponibles :

- La population étrangère et les immigrés étrangers
- Les logements de moins de 40 m^2 et de plus de 100m^2
- La population « hors ménage »
- Le revenu médian

Etant donné cette légère différence de sélection de variables de départ, on a appliqué trois fois la procédure de création d'indices socio-économiques précédemment décrite :

- en analyse « directe » sur les IRIS, similaire à celle présentée dans la section 7.3 ;
- en analyse inter-classes à l'échelle communale ;
- en analyse intra-classes à l'échelle des IRIS centrés par communes.

On présentera les résultats de ces trois procédures simultanément.

La première sélection de variables, destinée à choisir les meilleures représentantes des groupes de variables redondantes, est similaire entre les trois analyses (et celles déjà présentées précédemment) : il s'agit du chômage total, de la population active totale et de la mobilité de moins de 5 ans.

Les sélections de variables réalisées à la deuxième étape, c'est-à-dire les variables qui constitueront l'indice socio-économique, sont présentées dans la table 9.1.

Les sélections sont très proches pour l'ensemble des analyses. Ceci semble indiquer que ce sont essentiellement les mêmes variables qui expriment le mieux la variabilité socio-économique aussi bien à l'échelle communale que de l'IRIS (que l'on considère l'ensemble des IRIS les uns par rapport aux autres comme dans l'analyse directe ou uniquement au sein de leur commune).

Une fois cette sélection réalisée, le premier facteur de l'ACP réalisée sur celle-ci oppose bien des variables de défaveur à des variables de faveur pour chacune des analyses (voir figure 9.1) et l'on obtient donc bien un indice socio-économique « inter-classes » et « intra-classes ».

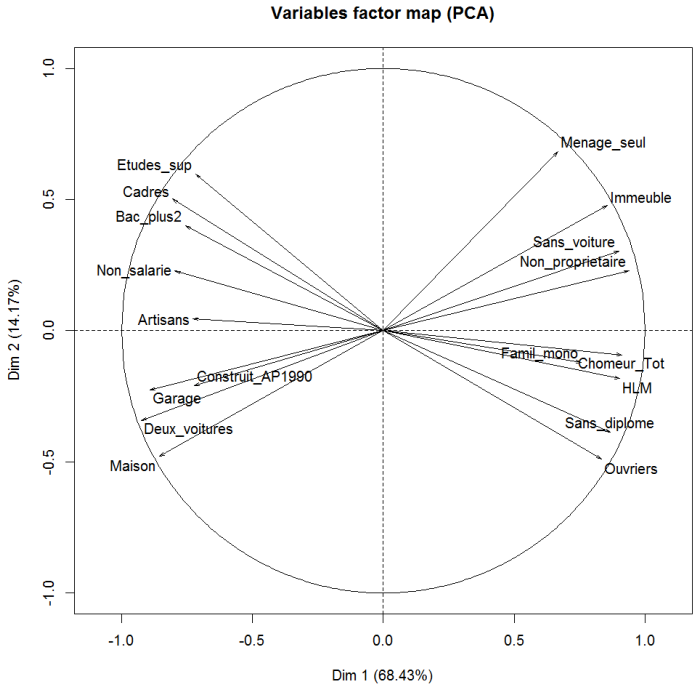
L'interprétation de ces indices est cependant différente.

De manière naturelle l'indice inter-classes est un indice entre communes qui place celles-ci suivant leur statut socio-économique relatif les unes aux autres. Il peut servir, à l'échelle de l'agglomération, pour déterminer les communes les plus favorisées et défavorisées mais aussi pour déterminer quelles sont les variables socio-économiques qui expliquent le plus les contrastes entre communes. Ainsi, dans le cas de l'agglomération de Lyon, la proportion de ménages n'étant pas propriétaires de leur logement, de ménages avec deux voitures ou plus ou encore le taux de chômage total font partie des variables qui contribuent le plus à l'indice socio-économique.

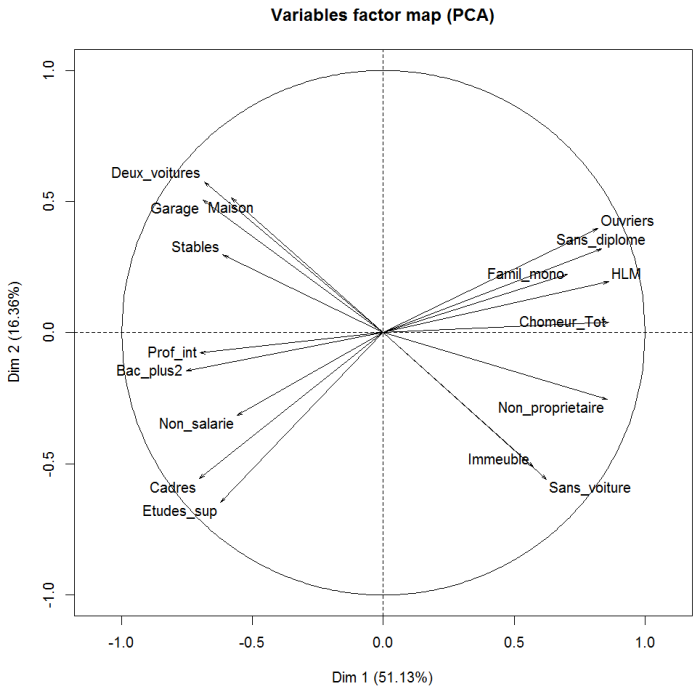
L'indice intra-classes paraît plus compliqué à interpréter. Dans la mesure où ils sont centrés par commune, il n'est réellement possible de comparer la position de deux IRIS que dans la mesure où ils sont dans la même commune. En effet, la composante communale étant retirée, la représentation de l'IRIS le plus favorisé d'une commune favorisée sera proche de celle de l'IRIS le plus favorisé d'une commune défavorisée. Sans précaution, ceci pourrait

TABLE 9.1 – Variables sélectionnées par les différentes analyses

	Analyse directe	**Analyse inter-classes**	**Analyse intra-classes**
Variables communes à toutes les analyses		Familles monoparentales Sans diplômes Bac +2 Chômage total Non salariés Cadres Ouvriers Maisons Immeubles Garages Non propriétaires HLM Sans voitures Deux voitures ou plus	
Autres variables	- Emplois stables Artisans Professions intermédiaires - -	Ménages composés d'une personne seule - Artisans - Etudes supérieures Construit après 1990	- Emplois stables - Professions intermédiaires Etudes supérieures -
Nombre de variables sélectionnées	17	18	17



(a) Analyse inter-classes



(b) Analyse intra-classes

FIGURE 9.1 – Cercles de corrélations des analyses inter et intra-classes

alors conduire à des interprétations erronées. On peut néanmoins remarquer que, dans le cas du Grand Lyon, certaines des variables qui contribuent le plus à expliquer la variation socio-économique entre IRIS, une fois retiré l'effet des communes, sont le chômage total, la proportion de HLM et la proportion de ménages non propriétaires de leur logement.

L'indice direct, enfin, a une interprétation similaire à celle réalisée dans les sections précédentes. Les IRIS de l'ensemble de l'agglomération sont alors positionnés les uns par rapport aux autres, et l'on peut identifier les variables qui expliquent le maximum de variance entre eux.

9.4 Discussion

Nous avons brièvement exposé l'une des techniques supplémentaires que nous avons explorées dans le cadre de cette thèse pour répondre à l'objectif de construction d'indicateurs composites.

Bien que la distinction entre analyse inter-classes et intra-classes puisse répondre de manière intéressante à différents buts (comparaison entre communes, comparaison de « l'importance » des différentes variables à des échelles différentes, ...), elle présente certaines difficultés d'interprétations et peut poser des « pièges » de communication (en particulier l'analyse intra-classes) qui limitent son application au contexte d'intérêt ici.

En effet, le but de l'application dans laquelle s'intègre ce travail est d'étudier les IRIS relativement les uns aux autres dans une agglomération entière, ce que ne permet pas cette approche. Par conséquent, seule l'approche « directe » a été employée et la réflexion sur cette technique d'analyse de données multi-niveaux n'a pas été développée davantage.

Conclusion et discussion de la deuxième partie

Dans cette partie, nous nous sommes concentrés sur l'objectif de construction d'indicateurs composites synthétisant d'importants jeux de données, en utilisant principalement pour cela les techniques d'analyses de données.

Le choix de ces techniques s'est rapidement imposé dans les différents cas de figures pour plusieurs raisons :

- D'abord parce que ce sont des techniques conçues et adaptées pour extraire l'information de bases de données de taille importante, qu'il semblait naturel d'utiliser d'un point de vue statistique.
- Ensuite, ces techniques permettent une approche « guidée par les données » que nous souhaitions avoir ici, par opposition aux sélections de variables « arbitraires » réalisées généralement dans le domaine d'application.
- Enfin, ce domaine d'application utilise encore peu ce type de méthodes alors qu'elles sont adaptées à plusieurs de ses problématiques.

Comme l'un des objectifs de cette thèse est d'étudier ce que peuvent apporter de « nouvelles » techniques par rapport à celles « traditionnellement » utilisées, nous avons choisi l'utilisation de l'analyse de données.

Dans les différents chapitres précédents, nous avons présenté et discuté le cheminement qui nous a conduit à développer une procédure de création d'indices socio-économiques, à construire un indice de multi-expositions environnementales et à étudier une méthode d'analyse de données multi-niveaux.

Il ressort de ces différents cheminements que, si la procédure de création d'indices socio-économiques présente de multiples avantages compensant la complexité de sa mise en œuvre (avantages qui ont conduit à l'utiliser dans de nombreux contextes), ce n'est pas encore entièrement le cas en l'état actuel des réflexions sur l'indice de multi-expositions environnementales. Plus précisément, ce dernier nous semble utile dans un contexte descriptif des différents profils d'exposition dans une zone mais doit encore être amélioré dans le cadre d'une utilisation pour mettre en évidence des effets sanitaires du cumul d'expositions.

De même, la méthode d'analyse de données multi-niveaux détaillée ici ne paraît pas adaptée aux buts de l'application de ce travail, d'autant que son interprétation complexe impose de prendre d'importantes précautions si l'on souhaite les communiquer efficacement.

Malgré tout, la mise au point d'indicateurs composites en utilisant l'analyse de données apparaît particulièrement intéressante dans le contexte de la santé publique :

- Elle permet de tirer parti des importantes bases de données mises à dispositions depuis ces dernières années d'une manière statistiquement efficace,
- Elle permet d'éviter d'introduire trop d'*a priori* ou d'arbitraire dans le choix des variables « d'intérêt »,
- Elle autorise des interprétations plus détaillées que de nombreuses autres méthodes généralement utilisées dans ce champ d'étude.

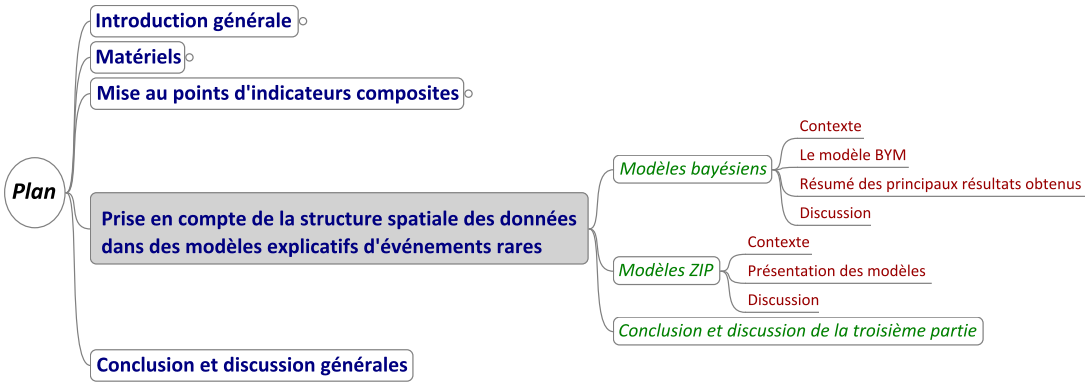
Troisième partie

Prise en compte de la structure spatiale des données dans des modèles explicatifs du risque d'événements rares

Prise en compte de la structure spatiale des données dans des modèles explicatifs du risque d'événements rares

Dans cette troisième partie nous détaillerons le deuxième partie principal de cette thèse : la prise en compte de la structure spatiale des données dans des modèles explicatifs du risque d'événements rares. L'un des principaux moyens que nous utiliserons pour cela sera les modèles bayésiens, et plus particulièrement le modèle « BYM » présenté au chapitre 10. Nous évoquerons également au chapitre 11 les modèles « ZIP » en termes de perspectives.

Dans chacun de ces chapitres il sera d'abord expliqué le contexte dans lequel le travail s'insère, puis les aspects méthodologiques et mathématiques seront détaillés, avant de résumer les principaux résultats obtenus et de discuter de l'ensemble. Lorsqu'une publication a été réalisée ou est en cours de soumission concernant ce travail, une présentation de celle-ci sera également insérée.



Chapitre 10

Modèles bayésiens

Dans ce chapitre nous détaillerons les modèles bayésiens utilisés dans le cadre de cette thèse. Nous expliciterons le contexte qui a mené à leur utilisation, avant de détailler le modèle « BYM » utilisé et de résumer les principaux résultats obtenus avec celui-ci. On notera que, pour l'ensemble de ce chapitre 10, les ouvrages de Christensen *et al.* [160], Robert [161], Ntzoufras [162], Lawson *et al.* [163] et Elliot *et al.* [164] ont été utilisés comme références, sans qu'ils soient systématiquement indiqués dans la suite.

10.1 Contexte

De très nombreuses études s'intéressant aux inégalités sociales et/ou environnementales de santé sont conduites à une échelle individuelle. Les données issues de ce type d'étude sont alors relativement faciles à analyser grâce à de nombreux modèles de régression « classiques » (linéaire, logistique, de Poisson, ...) adaptés à ces contextes. Cependant, certaines difficultés commencent à apparaître lorsque l'on intègre des données mesurées sur une zone plutôt que directement sur les individus. C'est notamment le cas en épidémiologie environnementale où, pour des raisons de coût et d'accessibilité des données, on dispose souvent uniquement d'une mesure des expositions sur une zone plutôt que d'expositions individuelles.

Jusqu'à récemment, dans ce type de situation, l'approche couramment adoptée était d'attribuer aux individus les mesures d'expositions dans leur zone puis de continuer d'utiliser (avec quelques précautions, comme la prise en compte du biais écologique) les modèles « classiques ». Cependant, des questions sont peu à peu apparues sur la possibilité d'un rôle de l'environnement (au sens large ici, incluant donc autant l'environnement physique que les caractéristiques démographiques et psycho-sociales) à la fois au niveau individuel et au niveau contextuel, en posant l'hypothèse que ces deux niveaux avaient des effets différents et simultanés sur la santé.

Par ailleurs, l'importance de prendre en compte la structure spatiale des données dans les modèles a été mise en évidence par différentes études (ceci a d'ailleurs été l'objet de l'un des premiers travaux conduit avant le début de cette thèse, qui sera davantage discuté par la suite et est présenté en annexe C.1)

Deux principaux types d'études se sont alors développés :

- D'une part les études « multi-niveaux » (qui ont été brièvement abordées dans le chapitre 9) prenant en compte simultanément ces deux niveaux ;

– D'autre part les études purement géographiques n'utilisant que des données contextuelles afin d'étudier cet effet en particulier.

> L'application de cette thèse s'inscrit dans ce deuxième type d'étude, puisque l'un des buts de ce projet est d'étudier la distribution géographique de mortalité infantile à l'échelle de l'IRIS sur les agglomérations retenues. Nous souhaitons par conséquent prendre en compte la structure spatiale de nos données, et en particulier dans le cas d'un modèle explicatif d'événement rare.

Ce type d'étude peut avoir une variété importante d'usages, mais l'on peut en distinguer deux qui nous intéressent [163, 164] :

– La **cartographie de mortalité** (*disease mapping*) vise à synthétiser les variations géographiques du risque de mortalité (ou de maladie) et à estimer les risques relatifs de mortalité sur une zone d'étude. Cela permet de répondre à un objectif descriptif (construction d'atlas), de fournir des informations en vue d'évaluer ou de mettre en place des politiques de santé ou encore de fournir des indices sur les liens entre ces risques et d'autres variables.

– L'**analyse écologique** (*ecological analysis* ou *geographical correlation study*) se concentre sur les liens entre les variations d'un événement de santé et celles de variables explicatives, toutes mesurées à une échelle géographique donnée (et non à l'échelle individuelle) et donc correspondant à des mesures « agrégées ». L'objectif est alors de déterminer ces liens pour expliquer l'événement de santé par les variables explicatives.

Bien que les buts de ces deux utilisations soient différents, les modèles statistiques employés pour celles-ci sont très proches et partagent de nombreux éléments communs. On retrouve notamment la plupart des difficultés de la cartographie de mortalité dans l'analyse écologique (celle-ci en ayant d'autres supplémentaires). Nous allons détailler dans la suite de cette section la progression de la modélisation conduisant des modèles classiques habituels à des modèles prenant en compte la structure spatiale des données, dans le cas d'événements rares.

Soit une zone d'étude Z composée de n unités géographiques (et statistiques). Sur cette zone d'étude, on compte le nombre d'occurrences d'un événement aléatoire par unité et on note y_i ce nombre pour l'unité i ($i = 1, \ldots, n$), réalisation d'une variable aléatoire Y_i. Dans notre cas, il s'agit donc d'une agglomération composée de n IRIS sur lesquels on compte le nombre de cas de mortalité infantile y_i.

On dispose également pour chaque unité des valeurs connues sur la période d'étude :

– De l'effectif de la **population « à risque »** n_i (ici le nombre de naissances vivantes). n_i est aléatoire.

– Du nombre de **cas « attendus »** e_i calculé sous l'hypothèse que le même taux d'occurrence de référence ρ s'applique sur toute la zone d'étude. On a alors $e_i = \rho n_i \ \forall i = 1, \ldots, n$. Il existe différentes manières de définir ρ : cela peut être un taux de référence sur une zone plus vaste (la France entière par exemple) ou bien un taux calculé à partir des données.

Dans notre cas, on pose $\rho = \frac{\sum_{i \in Z} y_i}{\sum_{i \in Z} n_i}$ c'est-à-dire le taux de mortalité infantile global de la zone. Donc e_i est aléatoire ici. Ce choix a été fait afin de ne pas dépendre de données autres que celles que l'on avait recueillies ainsi que, dans la mesure où l'on sait qu'il existe des diversités entre agglomérations, pour avoir un taux de mortalité par agglomération et comparer les taux des IRIS par rapport à celui-ci.

- Des valeurs $x_i^1, \ldots, x_i^p$ de **variables explicatives** $x^1, \ldots, x^p$ mesurées pour l'unité i.

On suppose également qu'il existe pour chaque unité un risque relatif θ_i d'occurrence de l'événement par rapport à la référence, *que l'on considérera dans un premier temps inconnu mais fixe* et que l'on va chercher à estimer.

Enfin, on notera respectivement $\underline{y}, \underline{e}, \underline{x}^j$ $(j = 1, \ldots, p)$ et $\underline{\theta}$ les vecteurs colonnes des y_i, e_i, x_i^j et θ_i ; $\underline{x}_i$ $(i = 1, \ldots, n)$ le vecteur ligne des x_i^j ; et X la matrice $n \times p$ des valeurs des variables explicatives mesurées sur les individus.

Une approche courante [164] pour estimer les risques relatifs d'événements rares consiste à employer un modèle de Poisson. On suppose alors que les cas observés sont indépendants et suivent une loi de Poisson : $Y_i | e_i \sim_{ind} \mathcal{P}(e_i \theta_i)$.

On montre que l'estimateur du maximum de vraisemblance de θ_i est $\hat{\Theta}_i = \frac{Y_i}{e_i}$. L'estimation $\hat{\theta}_i = \frac{y_i}{e_i}$ correspond à l'indicateur épidémiologique nommé ratio de mortalité standardisé (SMR, pour *standardised mortality ratio*), qui est couramment employé dans de nombreux contextes (n'étant pas forcément liés à la cartographie du risque) dès qu'il est nécessaire de comparer des taux de mortalité entre populations aux caractéristiques différentes.

Lorsque l'événement n'est pas « trop » rare, cet estimateur donne généralement une bonne illustration de la distribution géographique des risques. Cependant, pour des événements très rares et/ou de petites unités géographiques comme dans notre cas, les cartes représentant ces estimations sont compliquées à interpréter car les zones ressortant le plus sont souvent celles dont les données sont les moins fiables.

En effet, $\mathrm{Var}(\hat{\Theta}_i | e_i) = \frac{\mathrm{Var}(Y_i)}{e_i^2} = \frac{e_i \theta_i}{e_i^2} = \frac{\theta_i}{e_i}$ et cette variance peut être estimée par $\frac{y_i}{e_i^2}$. C'est donc pour les zones où les cas attendus sont les plus faibles (ce qui arrive d'autant plus fréquemment que l'événement est rare et les zones petites) que l'estimation du risque relatif a des chances d'être la plus importante mais aussi la moins précise.

Parallèlement, le nombre de cas observés dans des petites unités géographiques a souvent une variance plus importante que celle supposée par la distribution de Poisson. On a alors une variabilité dite « extra-poissonnienne » qui s'ajoute à celle de la distribution.

Une première possibilité pour estimer plus correctement les risques relatifs est d'inclure différents effets fixes de variables explicatives de la moyenne de la distribution de Poisson (et donc du risque relatif). On peut utiliser un modèle linéaire $\theta_i = \beta_0 + \underline{x}_i' \underline{\beta}$ mais le terme de droite peut alors parcourir l'ensemble des réels tandis que θ_i doit etre positif. Une solution est d'utiliser un modèle log-linéaire $\ln(\theta_i) = \beta_0 + \underline{x}_i' \underline{\beta}$ où l'on cherche à estimer :

- β_0, paramètre réel correspondant au log-risque relatif global dans la zone d'étude par rapport au taux de référence ;
- $\underline{\beta}$, vecteur $p \times 1$ de paramètres réels où le paramètre β_j associé à la variable explicative x^j correspond :
  - si x^j est quantitative, à l'augmentation du log-risque relatif pour une unité d'augmentation de x^j.
  - si x^j est l'indicatrice d'une modalité de variable qualitative, au log-risque relatif d'occurence de l'événement dans cette modalité j par rapport à la modalité de référence.

Cependant, ce type de modèle n'agit pas sur la variabilité extra-poissonnienne pouvant exister dans les données observées. *Une deuxième approche est alors de considérer que θ_i est aléatoire* et d'inclure dans le modèle des effets aléatoires supplémentaires visant à expliquer cette variabilité extra-poissonnienne. On se trouve donc dans le cadre de modèles linéaires généralisés à effets mixtes.

Dans les applications spatiales, ces effets peuvent être de deux types [163, 164] :
- d'une part une variabilité sans motif spatial (appelée généralement hétérogénéité non structurée), c'est à dire sans corrélation entre unités géographiques, que l'on peut assimiler aux résidus habituels d'analyse de régression ;
- d'autre part une variabilité avec une dépendance spatiale (appelée généralement hétérogénéité structurée), pour laquelle on suppose qu'il y a une corrélation entre unités voisines. Ceci peut arriver, par exemple, lorsqu'un facteur de risque inconnu ou non mesuré est lui-même distribué suivant un motif spatial.

Les modèles peuvent alors inclure l'un ou l'autre de ces effets, voire les deux. De ce choix dépendra le niveau de lissage global et local des risques relatifs effectué sur la carte. Dans ce travail, nous avons choisi d'utiliser un modèle combinant ces deux types d'hétérogénéité dans un cadre bayésien (dont on rappelle l'approche et la mise en œuvre générale en annexe A.2) : le modèle « BYM ».

10.2 Le modèle BYM

Dans la mesure où nous sommes dans un cadre bayésien et comme la loi de probabilités des données observées est modélisée par une loi de Poisson de moyenne $e_i\theta_i$, il est nécessaire de fournir une loi de probabilité *a priori* du vecteur $\underline{\theta}$. En effet, on rappelle que dans un cadre bayésien, on considère que l'on a des données observées $\underline{y}$ issues du vecteur aléatoire $\underline{Y}$ suivant une loi d'échantillonnage de paramètre $\underline{\theta}$ et de densité $f(\underline{Y}|\underline{\theta})$, d'où la vraisemblance $f(\underline{y}|\underline{\theta})$. On donne alors au paramètre $\underline{\theta}$ une loi *a priori* de densité $f(\underline{\theta})$ et, à partir de la vraisemblance et de cette loi *a priori*, on peut alors obtenir en utilisant le théorème de Bayes la loi *a posteriori* $f(\underline{\theta}|\underline{y})$ de $\underline{\theta}$.

Ainsi que nous l'avons énoncé précédemment, on suppose que les log-risques relatifs suivent un modèle linéaire $\ln(\theta_i) = \beta_0 + \underline{x}_i'\underline{\beta}$. On notera ici que, conformément au cadre bayésien, les paramètres β sont considérés comme des variables aléatoires ayant elles-mêmes une loi de probabilité *a priori* (et ceci nous fournit par conséquent la loi *a priori* de θ_i).

On va cependant ajouter à ce premier modèle des effets aléatoires permettant de prendre en compte la variabilité extra-poissonnienne des données [164].

Si l'on suppose que cette variabilité s'exprime sans motif spatial dans la zone, c'est-à-dire qu'il y a une hétérogénéité non structurée, cela revient à considérer que l'effet aléatoire associé est issu de variables aléatoires indépendantes. Une possibilité communément utilisée est alors d'ajouter au modèle un terme V_i : $\ln(\theta_i) = \beta_0 + \underline{x}_i'\underline{\beta} + V_i$, en supposant que le vecteur $\underline{V} = (V_1, \ldots, V_n)'$ suit une loi normale multivariée $\underline{V} \sim \mathcal{N}(0, \sigma_V^2 I)$ où σ_V^2 est la variance commune des composantes et I la matrice identité.

Si l'on suppose que la variabilité extra-poissonnienne a une structure spatiale, c'est-à-dire s'il y a une hétérogénéité structurée, on peut procéder de même et ajouter un terme U_i au modèle initial : $\ln(\theta_i) = \beta_0 + \underline{x}_i'\underline{\beta} + U_i$. Cependant, contrairement à l'hétérogénéité non structurée, cette dépendance spatiale peut être modélisée de nombreuses manières puisqu'elle est explicitement conçue pour refléter une ressemblance entre unités proches : il est ainsi possible de spécifier soit une distribution jointe au vecteur $\underline{U} = (U_1, \ldots, U_n)'$, soit des distributions conditionnelles univariées aux $U_i|U_{-i}$ $i = 1, \ldots, n$ (où $U_{-i} = \{U_j, j \neq i\}$).

Le modèle « de convolution » proposé par Besag, York et Mollié en 1991 [165] (plus connu sous le nom de modèle « BYM » en pratique, des initiales des noms des auteurs) combine des termes d'hétérogénéité non structurée et structurée indépendants dans un modèle hiérarchique (dont on rappelle le principe en annexe A.2.4). Pour ces auteurs, mettre simultanément ces deux termes permet de laisser les données décider de la proportion de risque résiduel due à une variation structurée et de la proportion due à une variation non structurée et c'est en particulier pour cette raison que nous avons choisi ce modèle.

> En effet, cela nous intéresse ici car, bien que l'on fasse l'hypothèse que les caractéristiques socio-économiques et/ou environnementales considérées dans ce travail ont une influence sur la mortalité infantile, il nous paraît par ailleurs évident que d'autres facteurs de risques, inconnus et non mesurés, structurés spatialement ou non, peuvent également entrer en jeu. Ainsi, certaines expositions environnementales non prises en compte ici pourraient avoir un effet sans suivre de structure spatiale (la pollution de l'air intérieur par exemple, ou encore certains composés ingérés) tandis que d'autres pourraient être très dépendants de ce point de vue (d'autres polluants de l'air que le NO_2, par exemple). Il s'agit par ailleurs d'un modèle qui est régulièrement employé dans des applications similaires à la nôtre [33, 163, 164].

Besag *et al.* proposent pour le terme d'hétérogénéité structurée l'utilisation d'une distribution Gaussienne intrinsèque autorégressive (*intrinsic Gaussian autoregressive*) définie par : $U_i|U_{-i} \sim \mathcal{N}(\bar{u}_i, \frac{\sigma_U^2}{m_i})$ avec $\bar{u}_i = \frac{1}{m_i}\sum_{j\in\delta_i} u_j$ où δ_i représente l'ensemble des voisins de i ; m_i le nombre de voisins de i ; et σ_U^2 est un paramètre de variance.

Un grand nombre de définitions du voisinage existe, mais la plus courante consiste à considérer que deux zones sont voisines si elles partagent une frontière commune (quelle que soit la longueur de cette frontière). Nous discuterons davantage de cette définition dans la section 10.5. Le choix d'une distribution conditionnelle est justifié par les auteurs par le fait que la modélisation est alors plus simple qu'avec une distribution jointe dans le cas de problèmes spatiaux. Intuitivement, on comprend bien que le but d'une telle distribution

est de s'appuyer sur les valeurs voisines d'autant plus fortement (avec une variance plus faible) que le nombre de voisins est important.

Pour finir, les distributions *a priori* sur les paramètres β_j $j = 0, \ldots, p$, σ_U^2 et σ_V^2 sont alors choisies en fonction des connaissances *a priori* ou en utilisant des distributions de référence. De manière générale, on choisit souvent pour les paramètres β_j des lois normales centrées de variance importante lorsqu'il n'y a pas d'*a priori* particulier. Concernant les paramètres σ_U^2 et σ_V^2, un choix courant (il s'agit de la loi conjuguée à la loi normale pour un paramètre de variance) est de modéliser leur inverse par une loi $\Gamma(\varepsilon, \varepsilon)$ avec ε très petit (0.01 ou 0.001). On notera que les interprétations de ces deux paramètres de variance sont différentes, le terme σ_V^2 déterminant une variabilité marginale tandis que le terme σ_U^2 détermine une variabilité conditionnelle (entre zones voisines).

> Nous avons choisi ici de conserver ces lois *a priori* de référence dans nos applications, dans la mesure où nous ne souhaitions pas introduire d'information *a priori*, celle-ci n'étant par ailleurs pas disponible dans notre cas. De plus, ces famille de lois « de référence » (définies précisément pour le cas de figure où l'on ne dispose pas d'information *a priori*) sont à notre connaissance les seules utilisées dans la littérature de ce domaine d'applications (avec peu de variétés dans leurs paramètres).

Nous avons par ailleurs réalisé des analyses de sensibilité sur les paramètres des lois gamma dans le cadre du travail présenté en annexe C.1 dont nous discuterons, ainsi que du choix des lois *a priori* en général, dans la section 10.5.

Pour conclure cette présentation du modèle BYM, on peut donc synthétiser les étages qui le composent (modèle sur les données, modèle sur les paramètres $\underline{\theta}$, lois *a priori*) de la manière suivante :

$$\forall i = 1, \ldots, n \begin{cases} Y_i | e_i, \theta_i \sim_{ind} \mathcal{P}(e_i \theta_i) \\ \ln(\theta_i) = \beta_0 + \underline{x}_i' \underline{\beta} + U_i + V_i \\ V_i \sim_{i.i.d} \mathcal{N}(0, \sigma_V^2) \\ U_i | U_{-i} \sim \mathcal{N}(\bar{u}_i, \frac{\sigma_U^2}{m_i}) \\ \beta_j \sim_{i.i.d} \mathcal{N}(0, 10^4) \; j = 0, \ldots, p \\ \frac{1}{\sigma_U^2} \sim \Gamma(0.01, 0.01) \\ \frac{1}{\sigma_V^2} \sim \Gamma(0.01, 0.01) \end{cases}$$

10.3 Résumé des principaux résultats obtenus

Pour des raisons de disponibilité des données sanitaires, le modèle « BYM » a été appliqué à Paris intra-muros, au Grand Lyon et à Lille Métropole. Les cas de mortalité infantile ou néonatale ont été modélisés en utilisant différents ensembles de variables explicatives :

- sans variables explicatives (uniquement avec les termes d'hétérogénéité structurée et non structurée) ;
- avec l'indice socio-économique défini dans la section 7.2 sous forme quantitative ;
- avec l'indice socio-économique en classes.

Dans le cas de l'agglomération de Lyon, à ces modèles s'ajoutent également ceux avec les variables explicatives suivantes :

- avec l'indice de multi-expositions environnementales défini dans la section 8 ;
- avec l'indice de multi-expositions environnementales et l'indice socio-économique quantitatif ;
- avec l'indice de multi-expositions et l'indice socio-économique quantitatif ainsi que leur interaction.

Des modèles de Poisson (sans termes spatiaux) ont également été testés avec les mêmes variables explicatives sur les mêmes données afin de comparer leurs résultats et ceux obtenus avec le modèle « BYM ». Nous exposerons ici une synthèse des résultats obtenus pour ces différents modèles dans le cas de la mortalité infantile.

Les lois *a posteriori* ont été estimées à l'aide du logiciel WinBUGS [166]. Ce logiciel utilise des méthodes de Monte-Carlo par chaînes de Markov et l'échantillonneur de Gibbs (dont les principes sont rappelés en annexe A.2.5) pour évaluer les différentes distributions *a priori*. C'est en particulier le développement des techniques de simulation numérique permettant de simuler des échantillons de lois de probabilité et d'estimer des intégrales compliquées qui a permis d'utiliser davantage les techniques bayésiennes et de baser le choix des lois *a priori* soit sur des connaissances passées ou l'avis d'un expert, soit sur des distributions *a priori* de référence, plutôt que sur des considérations purement mathématiques de facilité d'analyse.

Nous avons itéré deux chaînes de Markov durant 60 000 itérations (dont 30 000 de rodage, ce qui semble très largement suffisant pour atteindre la convergence). La convergence de l'algorithme a été vérifiée à l'aide des graphiques des chaînes, de l'autocorrélogramme et du diagnostic de Gelman-Rubin (rappelés en annexe A.2.6). On obtient donc 30 000 simulations suivant les lois *a posteriori* permettant d'approcher celles-ci.

Une fois les distributions *a posteriori* des paramètres obtenues, on choisit comme indicateurs de synthèse de celles-ci la moyenne *a posteriori* (la médiane ou le mode *a posteriori* sont aussi utilisés dans la littérature). Dans le cas des paramètres β_j, on utilise également l'intervalle crédible (IC) à 95% comme indicateur de dispersion. Cet intervalle est défini simplement comme l'intervalle entre le percentile 2,5% et le percentile 97,5% de la distribution *a posteriori*, qui contient donc 95% des valeurs de la distribution *a posteriori*. C'est l'interprétation de cet intervalle (suivant qu'il contient 0 ou non) qui nous conduira à dire qu'un paramètre est significativement non nul, par analogie avec l'interprétation des intervalles de confiance de l'approche classique.

Pour faciliter l'interprétation des coefficients du modèle linéaire, on présentera leurs exponentielles. En effet, ainsi qu'on l'a vu en section 10.1 on peut interpréter le coefficient $\exp(\beta_j)$ comme le coefficient multiplicateur du risque relatif de mortalité pour une unité d'augmentation de x^j (dans le cas d'une variable quantitative), ou comme le risque relatif de mortalité dans la modalité dont j est l'indicatrice par rapport à la modalité de référence (dans le cas d'indicatrices de modalités d'une variable qualitative). La valeur de référence pour interpréter l'intervalle crédible sera alors 1.

La table 10.1 résume les résultats obtenus pour les trois agglomérations en appliquant les modèles avec effets spatiaux sans variables explicatives, puis avec comme variables explicatives l'indice socio-économique (sur les données 2006) en quantitatif, et l'indice socio-économique en trois classes (basées sur les tiertiles).

TABLE 10.1 – Moyenne et intervalle crédible à 95% des distributions *a posteriori* des paramètres explicatifs de mortalité infantile pour différents modèles

Modèle		**Paris (intra-muros)**	**Lyon**	**Lille**
Effets spatiaux uniquement	Constante	0,86 (0,77-0,96)	0,94 (0,86-1,03)	0,90 (0,81-0,99)
Indice SES quantitatif et effets spatiaux	Constante	0,86 (0,77-0,95)	0,92 (0,84-1,00)	0,87 (0,79-0,96)
	Indice SES	1,20 (1,11-1,31)	1,21 (1,11-1,31)	1,29 (1,19-1,40)
Indice SES en trois classes et effets spatiaux	Constante	0,78 (0,64-0,94)	0,74 (0,62-0,88)	0,64 (0,52-0,77)
	Classe 1[a]	1,00	1,00	1,00
	Classe 2	0,98 (0,77-1,25)	1,16 (0,92-1,46)	1,29 (1,01-1,65)
	Classe 3[b]	1,35 (1,06-1,71)	1,61 (1,30-2,00)	1,98 (1,58-2,49)

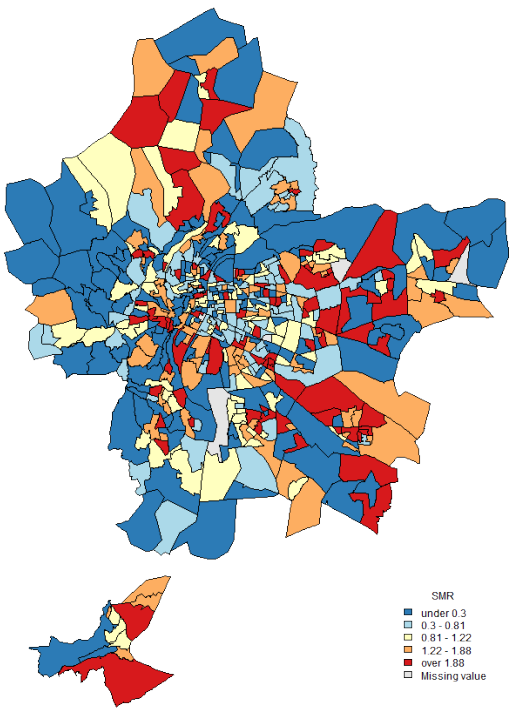
Pour des raisons de place, seul le nom des villes principales des agglomérations a été indiqué.
[a] classe des plus favorisés (référence) ; [b] classe des plus défavorisés

On remarque que l'indice socio-économique a un effet significatif sur la mortalité infantile pour les trois agglomérations. Ainsi, une augmentation d'une unité d'indice socio-économique dans un quartier (donc une augmentation de sa défaveur) augmenterait de 20% à 29% le risque de mortalité infantile dans ce quartier. Lorsque l'on étudie le statut socio-économique en trois classes en prenant comme classe de référence les 33% d'IRIS les plus favorisés, on remarque que l'effet est particulièrement accru pour les IRIS les plus défavorisés. Ainsi, le risque de mortalité infantile dans les 33% d'IRIS les plus défavorisés est 1,35 à 1,98 fois plus élevé que dans les 33% d'IRIS les plus aisés.

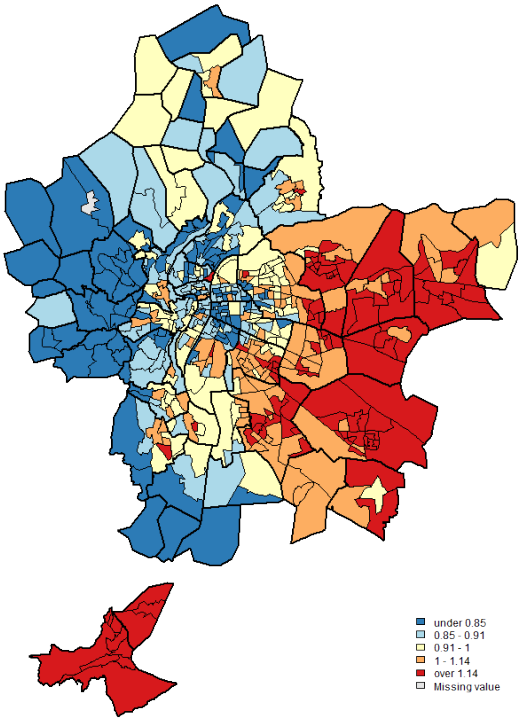
En prenant l'exemple du Grand Lyon, on constate en observant la carte des SMR (figure 10.1a) qu'il existe une importante variabilité spatiale des estimations du risque de mortalité infantile en utilisant cette méthode. Ainsi, des IRIS présentant un SMR de plus de 1,88 peuvent être entièrement entourées d'IRIS avec une estimation du risque inférieur à 0,3. Partant du principe que des IRIS voisins partagent des caractéristiques voisines, il est alors possible de s'interroger sur la fiabilité de telles estimations.

En appliquant le modèle BYM sans variables explicatives (et donc uniquement avec les termes d'hétérogénéité structurée et non structurée, on peut alors prendre en compte la structure spatiale des données dans les estimations et effectuer un « lissage » (figure 10.1b). Ceci permet à la fois de consolider les estimations en tirant parti de la ressemblance de caractéristiques entre unités voisines et de prendre en compte les facteurs de risques (structurés ou non) potentiellement non mesurés. On remarque alors que les risques de mortalités sont estimés comme plus importants dans l'est de l'agglomération.

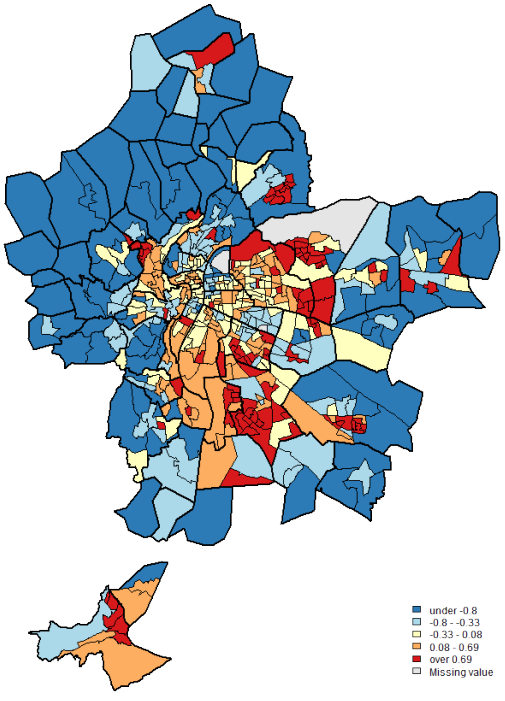
On peut rapprocher ceci du statut socio-économique (représenté par l'indice, figure 10.1c), qui est également plus faible dans cette partie de l'agglomération. L'ajustement du modèle par le statut socio-économique permet de fournir d'autres estimations des risques, qui demeurent importants à l'est de Lyon (figure 10.1d), et de montrer les liens entre cette variable explicative et la mortalité infantile (table 10.1).



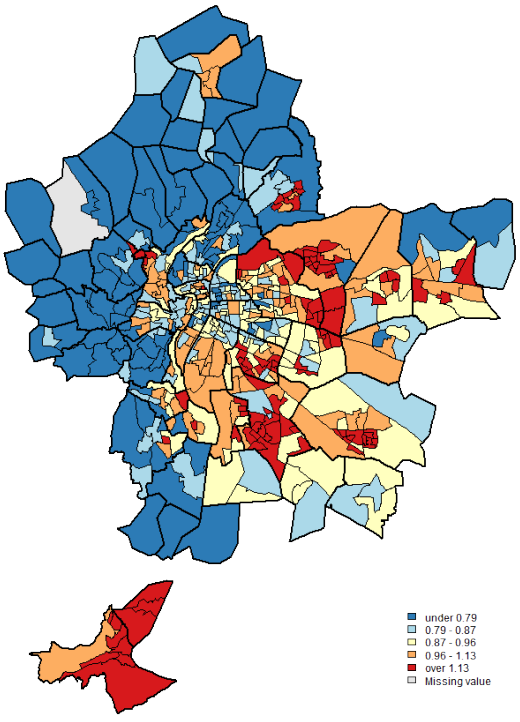
(a) Estimation par SMR



(b) Modèle BYM avec termes spatiaux



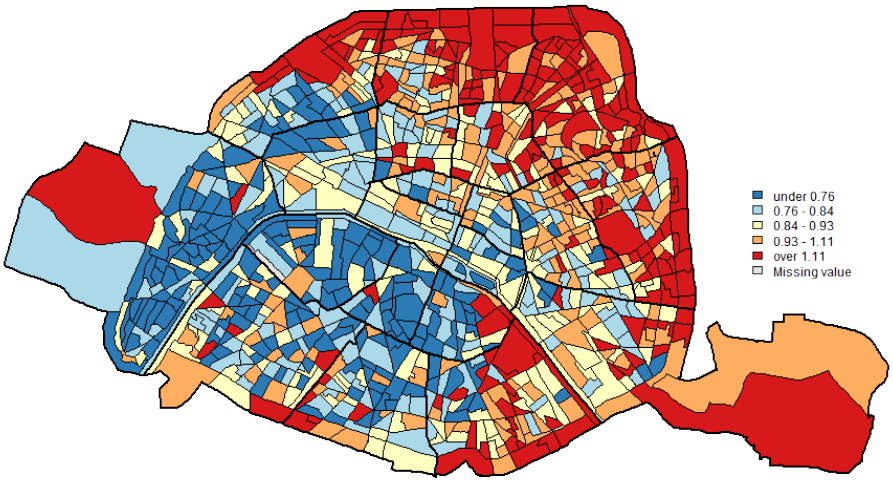
(c) Indice socio-économique



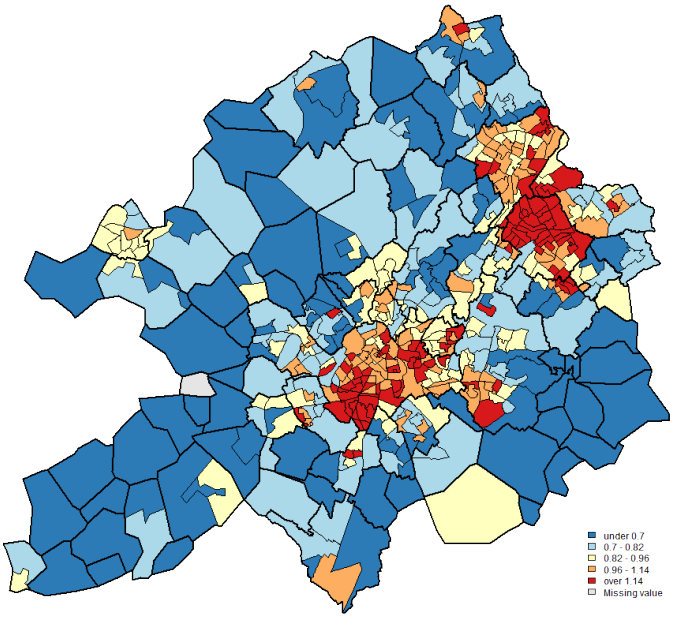
(d) Modèle BYM avec SES quantitatif et termes spatiaux

FIGURE 10.1 – Estimations des risques relatifs suivant différents modèles pour le Grand Lyon

Après ajustement sur l'indice socio-économique, les zones de plus fort risque relatif dans le Grand Lyon se concentrent essentiellement à l'est et au sud-est de la commune de Lyon proprement dite. Dans le cas de Paris intra-muros (figure 10.2a), ce sont davantage les IRIS du nord-est et de l'est qui ont des estimations plus élevées ainsi que ceux proches du périphérique. Enfin, pour Lille Métropole (figure 10.2b), ce sont particulièrement les villes de Lille, Tourcoing et Roubaix et leurs proches alentours qui regroupent les plus fortes estimations. On voit donc que les motifs spatiaux du risque de mortalité infantile sont très différents suivant les agglomérations.



(a) Paris intra muros



(b) Lille Métropole

FIGURE 10.2 – Estimations des risques relatifs par le modèle BYM avec SES quantitatif et termes spatiaux pour Paris intra muros et Lille Métropole

N.B. : les échelles géographiques sont différentes entre les cartes.

La table 10.2 présente une comparaison des estimations des paramètres des variables explicatives suivant que l'on utilise simplement un modèle de Poisson log-normal (sans effets d'hétérogénéité spatiale structurée et non structurée) ou le modèle BYM. Afin d'étudier la qualité d'ajustement des modèles, on utilise le DIC (*Deviance Information Criterion*)[167], qui est un équivalent bayésien du critère d'Akaike (AIC).

Si on note la déviance du modèle $D(\underline{\theta}) = -2\log(f(\underline{y}|\underline{\theta}))$; $\overline{D(\underline{\theta})}$ la moyenne *a posteriori* de la déviance et $D(\bar{\theta})$ une estimation de la déviance à partir de la moyenne *a posteriori* de $\underline{\theta}$; alors $\text{DIC} = 2\overline{D(\underline{\theta})} - D(\bar{\theta}) = D(\bar{\theta}) + 2p_D$ où $p_D = \overline{D(\underline{\theta})} - D(\bar{\theta})$ est un terme visant à évaluer la complexité du modèle et est appelé nombre de paramètres effectifs par analogie avec l'AIC (dans le cas de modèles non-hiérarchiques, ce nombre est environ égal au nombre réel de paramètres). Le DIC s'interprète de la même manière que l'AIC, c'est-à-dire que plus il est faible, meilleur est le modèle.

Bien que les estimations des $\exp(\beta_j)$ elles-mêmes soient très légèrement plus élevées lorsqu'il n'y a pas d'effets spatiaux, le DIC des modèles avec effets spatiaux indique qu'ils ont une meilleure qualité d'ajustement (ces deux constats sont faits pour les trois agglomérations). Cette meilleure qualité d'ajustement provient certainement du fait que les effets spatiaux ajoutés permettent de prendre en compte d'autres variables explicatives inconnues ou non mesurées, et que cette prise en compte est suffisamment importante pour justifier l'ajout de deux paramètres au modèle.

TABLE 10.2 – Comparaison entre modèles avec et sans effets spatiaux pour Lille Métropole.

Modèle	**Variables**	**Avec effets spatiaux**		**Sans effets spatiaux**	
		Estimations	DIC[c]	Estimations	DIC[c]
Constante seule	Constante	0,90 (0,81-0,99)	1 402,09	1,00 (0,93-1,08)	1 438,01
Indice SES quantitatif	Constante	0,87 (0,19-0,96)	1 384,37	0,91 (0,84-1,39)	1 388,09
	Indice SES	1,29 (1,19-1,40)		1,30 (1,21-1,39)	
Indice SES en trois classes	Constante	0,64 (0,52-0,77)	1 383,54	0,66 (0,55-0,78)	1 385,71
	Classe 1[a]	1,00		1,00	
	Classe 2	1,29 (1,01-1,65)		1,30 (1,03-1,64)	
	Classe 3[b]	1,98 (1,58-2,49)		2,03 (1,65-2,50)	

Moyennes et intervalles crédibles à 95% des distributions *a posteriori* des paramètres explicatifs de mortalité infantile
[a]classe des plus favorisés (référence) ; [b]classe des plus défavorisés ; [c]Deviance Information Criterion

Enfin, la table 10.3 présente les résultats pour le Grand Lyon des modèles incluant l'indice de multi-expositions environnementales comme variable explicative en prenant la classe des IRIS « intermédiaires » (IRIS de zones pavillonnaires plus proches des centres-villes mais moins exposés au bruit, à la pollution de l'air ou au trafic que la moyenne ; voir section 8.3).

On remarque que l'indice de multi-expositions ne semble pas avoir d'influence significative sur la mortalité infantile, qu'il soit utilisé seul ou combiné à l'indice socio-économique. Comme nous en avons discuté en section 8.5, ce résultat semble contredire les liens qui ont pu être montrés (avec d'autres types de modèles) entre les expositions environnementales considérées individuellement et la mortalité infantile. Ainsi que nous l'avons détaillé en

TABLE 10.3 – Moyenne et intervalle crédible à 95% des distributions *a posteriori* des paramètres explicatifs de mortalité infantile pour différents modèles sur le Grand Lyon

Modèle	**Paramètre de la variable**	**Estimations**
Indice de multi-exposition et effets spatiaux	Constante	1,05 (0,89-1,24)
	Classe 1 (IRIS « divers »)	1,93 (0,30-7,56)
	Classe 2 (IRIS « ruraux »)	0,69 (0,42-1,11)
	Classe 3(IRIS « industries »)	0,93 (0,72-1,20)
	Classe 4 (IRIS « intermédiaires »)	1,00
	Classe 5 (IRIS « centre-ville »)	0,84 (0,68-1,05)
Indices socio-économique, de multi-exposition et effets spatiaux	Constante	1,03 (0,88-1,20)
	Classe 1 (IRIS « divers »)	1,71 (0,26-6,62)
	Classe 2 (IRIS « ruraux »)	0,84 (0,52-1,34)
	Classe 3(IRIS « industries »)	0,85 (0,66-1,09)
	Classe 4 (IRIS « intermédiaires »)	1,00
	Classe 5 (IRIS « centre-ville »)	0,83 (0,68-1,02)
	Indice SES	1,21 (1,11-1,32)
Indices socio-économique, de multi-exposition, interactions et effets spatiaux	Constante	1,03 (0,88-1,21)
	Classe 1 (IRIS « divers »)	0,05 (0,00-3,73)
	Classe 2 (IRIS « ruraux »)	0,76 (0,29-1,75)
	Classe 3(IRIS « industries »)	0,83 (0,61-1,13)
	Classe 4 (IRIS « intermédiaires »)	1,00
	Classe 5 (IRIS « centre-ville »)	0,83 (0,67-1,36)
	Indice SES	1,20 (1,06-1,36)
	Interaction classe 1 et SES	3,83 (0,82-118,04)
	Interaction classe 2 et SES	0,94 (0,46-1,77)
	Interaction classe 3 et SES	1,03 (0,83-1,27)
	Interaction classe 4 et SES	1,00
	Interaction classe 5 et SES	1,00 (0,81-1,22)

section 8.5, ceci pourrait s'expliquer par le fait que l'indice de multi-expositions agrège trop les données et « gomme » l'effet des variables ou au fait que cet indice ne comporte pas assez d'expositions environnementales qui permettraient d'avoir une idée encore plus fine du « fardeau environnemental » pesant sur les IRIS.

10.4 Applications du modèle BYM dans d'autres travaux

Outre les résultats précédents, nous avons au cours de cette thèse appliqué le modèle « BYM », dans d'autres travaux et contextes.

L'influence des inégalités sociales et du sexe sur la survenue d'infarctus du myocarde à Strasbourg, à l'aide de modèles bayésiens (voir « *A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex : a Bayesian modeling approach* » en annexe C.1).

Dans ce travail préliminaire à cette thèse, nous avons d'une part appliqué le modèle BYM de manière similaire à celle exposée dans la section précédente pour expliquer les risques d'infarctus du myocarde en prenant comme variable explicative le statut socio-économique contextuel et en stratifiant les analyses sur le sexe.

D'autre part, nous avons également comparé les résultats du modèle BYM avec ceux de modèles n'incluant pas de termes spatiaux, n'incluant que l'hétérogénéité structurée ou non structurée, ou incluant les deux. Nous avons alors constaté l'importance d'ajouter les termes spatiaux pour la qualité d'ajustement du modèle et des changements (légers dans ce cas) dans les estimations des risques.

Enfin, dans ce même travail, nous avons également effectué des analyses de sensibilité entre trois ensembles de paramètres des lois gamma ((0.01, 0.01), (0.001, 0.001) et (0.5, 0.0005)) utilisés dans la littérature afin d'étudier l'influence de ceux-ci. Nous avons ainsi pu constater la faible influence (quelques variations pour certaines deuxièmes décimales) de ces paramètres sur les estimations finales, et nous avons choisi d'utiliser des lois $\Gamma(0.01, 0.01)$ (y compris dans les applications suivantes) car le modèle les utilisant présentait un DIC légèrement meilleur (1 342,57 contre 1 344,51 et 1 346,14).

Les inégalités sociales de mortalité infantile à Lille, dans un article que nous avons déjà présenté en section 7.6 et exposé en annexe C.2.

Dans ce travail, outre l'application de l'indice socio-économique que nous avons déjà décrite, nous avons également mis en œuvre le modèle BYM afin d'expliquer la mortalité infantile à l'aide de l'indice socio-économique de manière similaire à celle présentée dans la section précédente. Nous avons également dans ce travail effectué les différentes comparaisons (à l'aide du DIC) entre modèles avec et sans termes spatiaux

> Dans ces deux travaux additionnels comme dans les résultats précédemment exposés, en effectuant des comparaisons entre le modèle BYM et les modèles sans termes spatiaux, on remarque des changements, faibles mais présents, dans les coefficients et un meilleur DIC pour le modèle BYM.

10.5 Discussion

Dans ce chapitre, nous avons présenté le modèle « BYM » comme l'une des solutions possibles pour prendre en compte la structure spatiale des données dans un modèle explicatif d'événement rare. Dans un cadre bayésien, ce modèle ajoute au modèle de Poisson classique deux termes aléatoires représentant l'hétérogénéité non structurée et l'hétérogénéité structurée spatialement, ce qui conduit à effectuer un « lissage » des risques estimés pour tenir compte des ressemblances de caractéristiques entre IRIS voisins.

Ce modèle, qui a déjà été utilisé dans plusieurs études écologiques géographiques s'intéressant aux inégalités sociales et/ou environnementales de santé, permet ainsi de prendre en compte de manière efficace les données spatialisées. Cette prise en compte apparaît comme importante dans les analyses, puisque ne pas l'effectuer pourrait conduire à sur ou sous-estimer les risques de survenue des événements considérés, ainsi que nous avosn pu le mettre en évidence aussi bien dans les résultats exposés en section 10.3 que dans les autres applications présentées dans la section 10.4.

Parmi ses limitations, on peut citer le fait que le modèle BYM soit inscrit dans un cadre bayésien peu familier de la plupart des chercheurs en épidémiologie. Cette particularité peut causer des difficultés de mise en œuvre (concernant le choix des distributions *a priori* par exemple), d'interprétation des résultats (bien que le bayésien permette une grande variété de résultats grâce à la disponibilité de la distribution *a posteriori*, ce sont souvent des estimations ponctuelles et des intervalles crédibles qui sont présentés par analogie aux résultats d'analyses « statistiques ») et plus généralement de communication dans les champs d'application. En effet, dans la mesure où cette technique est moins connue et maîtrisée (en épidémiologie sociale et environnementale) que d'autres modèles « classiques », il est nécessaire de faire d'importants efforts pour expliquer cette méthode et justifier son intérêt (là où l'utilisation d'un modèle classique n'aurait posé aucun problème).

Ce cadre bayésien implique également que l'application du modèle BYM ne peut pas être effectuée avec tous les logiciels statistiques usuels. Il faut alors soit que l'utilisateur effectue des développements spécifiques supplémentaires, soit qu'il se serve de packages ou de logiciels dédiés (comme WinBUGS). Les techniques utilisées (MCMC) sont par ailleurs plus coûteuses en temps de calcul que les algorithmes d'estimation des modèles classiques, ce qui conduit parfois à devoir patienter durant un temps important (parfois plus d'une heure) avant d'obtenir un résultat (contrairement aux modèles classiques donnant des résultats presque immédiatement).

Le choix des distributions *a priori* fait également partie des difficultés potentielles d'utilisation. Très généralement dans la littérature, celles-ci sont choisies parmi les distributions *a priori* de référence. Il peut alors être nécessaire d'effectuer quelques analyses de sensibilité (ainsi que nous l'avons évoqué plus haut) afin de vérifier que les paramètres des lois *a priori* n'ont pas une influence trop importante sur les résultats (cela ne devrait pas être le cas si les données sont de taille suffisante).

Mais il est également possible de donner une loi *a priori* basée sur des connaissances préalables, d'autres données ou un avis d'expert, ce qui fait partie des avantages du bayésien.

Bien que nous n'ayons pas effectué ce type de travail faute de compétences en « *elicitation* » (le terme anglais utilisé pour définir ce processus de modélisation des connaissances d'un expert en une loi *a priori*), l'utilisation d'autres types de loi *a priori* dans le modèle BYM fait partie des perspectives d'extension de cette application.

La prise en compte de la structure spatiale dépend également de la définition du voisinage. Ici, nous avons choisi comme dans de nombreuses autres études de considérer que les voisins d'un IRIS étaient tous les IRIS adjacents à celui-ci. Cette définition ne pose pas de problème particulier lorsque les unités géographiques sont de tailles comparables et sont organisées suivant un motif régulier (quadrillage par exemple), mais peut amener quelques difficultés lorsque ce n'est pas le cas. Ainsi, dans le cas de nos zones d'études, nous disposons à la fois d'IRIS de centre-ville de faible superficie, et d'IRIS en périphérie, moins densément peuplés et donc de plus grande taille. En utilisant comme définition du voisinage le partage d'une bordure commune, la notion de distance est alors supprimée, ce qui implique alors que des IRIS géographiquement proches (et pouvant donc avoir des caractéristiques communes) pourraient ne pas être considérés comme voisins, par exemple dans le cas d'IRIS de centre-ville. Inversement, des IRIS de plus grande superficie peuvent être considérés comme voisins alors que leurs points extrêmes peuvent être géographiquement très éloignés (ceci est illustré figure 10.3 par le cas de l'agglomération de Marseille).

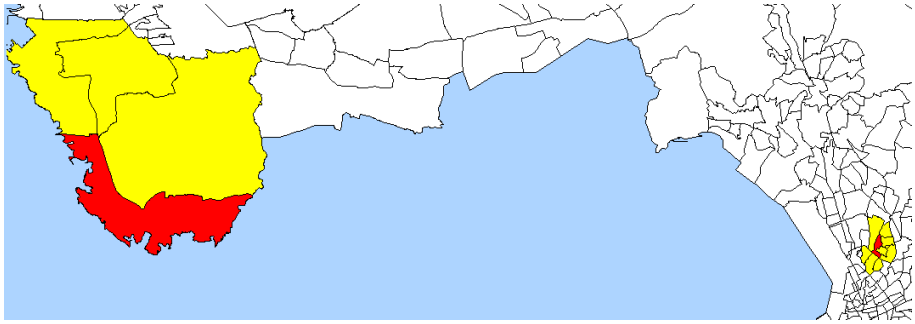


FIGURE 10.3 – Illustration de la définition du voisinage, agglomération de Marseille.

En rouge deux IRIS sélectionnés, en jaune leurs voisins adjacents

D'autres méthodes pour définir le voisinage sont alors possibles : on peut ainsi par exemple prendre les voisins de second ordre (en se basant toujours sur le partage de frontières, mais en incluant les « voisins des voisins »), ou encore se baser sur des critères de distance géographique. Cette dernière définition, si elle peut être plus pertinente, est par contre plus complexe à mettre en œuvre et c'est pourquoi elle n'a pas été employée ici. *La comparaison de différentes définitions du voisinage et l'étude de leur influence sur les résultats fait néanmoins partie des perspectives de ce travail.*

Enfin, la question du comportement du modèle BYM aux bords des zones n'a pas été évaluée ici. En effet, des « effets de bords » sont possibles ici puisque la zone d'étude est finie et a comme limite soit la mer (dans le cas de l'agglomération de Marseille par exemple), soit une frontière avec un autre pays (dans le cas de Lille Métropole), soit tout simplement la fin de la zone d'étude considérée. Dans ces deux derniers cas, les IRIS aux bords de la zone ont bien des voisins, mais ceux-ci (et les données les concernant) n'ont

pas été inclus et leurs effets ne sont pas pris en compte. Le modèle considère donc qu'il n'y a aucun voisin à cet endroit, ce qui peut par conséquent influencer le calcul du terme d'hétérogénéité spatiale structurée (puisque les paramètres de la loi *a priori* de ce terme dépendent notamment du nombre de voisins) et le lissage des risques qui en découle.

> Malgré ces différents points, le modèle BYM permet de prendre en compte la structure spatiale des données de manière efficace et a permis de mettre en évidence à plusieurs reprise et de façon consistante des liens entre le statut socio-économique et la mortalité infantile dans toutes les agglomérations de l'étude.

Chapitre 11

Les modèles ZIP

11.1 Contexte

On a vu dans le chapitre précédent que dans le cadre d'analyses spatiales, il était courant de supposer que les données (le nombre de cas observés par IRIS) suivaient une loi de Poisson mais qu'il était souvent nécessaire de prendre en compte une variabilité extra-poissonnienne des données. Nous avons présenté le modèle BYM comme une réponse à ce problème, mais de nombreuses autres solutions sont possibles. En particulier, lorsque la variabilité se traduit par un nombre plus important que prévu de 0, il est possible d'utiliser des modèles *Zero-inflated Poisson* (ZIP). Notons que l'on peut définir une version *zero-inflated* pour de nombreuses autres distributions (par exemple la loi binomiale ou la loi binomiale négative).

Ces modèles sont utilisés dans de nombreux domaines (santé publique, sociologie, psychologie, ingénierie, agriculture ...) lorsqu'un excès de 0 est constaté. Ils modélisent alors ce phénomène en ajoutant un nouveau paramètre de probabilité, permettant d'expliquer la génération des 0 excédentaires. Bien que ces modèles puissent être utilisés aussi bien en statistique classique que bayésienne, nous allons présenter ceux-ci avec l'approche bayésienne dans la continuité de ce qui a déjà été présenté.

11.2 Présentation des modèles

Afin de modéliser l'excès de zéros par rapport à la loi de Poisson, la distribution ZIP est constituée d'un mélange d'une loi de Dirac en 0 et d'une loi de Poisson de paramètre θ, $(1-\pi)\delta_0 + \pi\mathcal{P}(\theta)$, $\pi \in [0;1]$, de densité

$$\mathbb{P}(Y = y; \pi, \theta) = \begin{cases} (1-\pi) + \pi\exp(-\theta) & y = 0 \\ \pi\exp(-\theta)\frac{\theta^y}{y!} & y = 1, 2, \dots \end{cases}$$

qui a pour espérance $\mathbb{E}[Y; \pi, \theta] = \pi\theta$ et pour variance $\mathrm{Var}(Y; \pi, \theta) = \pi\theta(1 + (1-\pi)\theta)$. Contrairement à une loi de Poisson, la moyenne et la variance ne sont donc pas égales et la variance est plus importante que la moyenne, ce qui permet de prendre en compte la variabilité supplémentaire.

Si l'on se place dans le contexte précédemment présenté chapitre 10, on peut alors définir un modèle bayésien hiérarchique. En effet, on remarque tout d'abord qu'une loi de Poisson

de moyenne nulle est une loi de Dirac en 0. Simuler la distribution ZIP revient alors à simuler une loi de Poisson de paramètre λ où $\lambda = 0$ avec probabilité $(1 - \pi)$ et $\lambda = \theta$ avec probabilité π. Pour cela, on peut poser que $\lambda = \xi\theta$, où ξ prend la valeur 0 avec probabilité $(1 - \pi)$ et la valeur 1 avec probabilité π.

On peut également considérer, et c'est ce que l'on fera par la suite, que les paramètres λ, ξ, θ et π dépendent de l'IRIS, et on les notera donc λ_i, ξ_i, θ_i et π_i (notons qu'il est parfaitement possible de réaliser un modèle où le paramètre π serait commun à tous les IRIS, par exemple).

Dans ce cadre, il devient ensuite possible de modéliser la probabilité de succès π_i à l'aide d'un modèle logistique (ayant p_α variables explicatives) et le paramètre θ_i par un modèle log-normal comme précédemment (avec $p_\beta$variables explicatives). Les variables explicatives de ces deux modèles peuvent être les mêmes ou être propres à chacun. Enfin, on considérera que les paramètres de ces deux modèles suivent des lois *a priori* normales de référence.

On obtient alors la formulation suivante :

$$\forall i = 1, \ldots, n \begin{cases} Y_i|\lambda_i \sim_{ind} \mathcal{P}(\lambda_i) \\ \lambda_i = \xi_i\theta_i \\ \xi_i|\pi_i \sim_{ind} \mathcal{B}(1, \pi_i) \\ \text{logit}(\pi_i) = \alpha_0 + \underline{x}_i'\underline{\alpha} \\ \ln(\theta_i) = \beta_0 + \underline{x}_i'\underline{\beta} \\ \alpha_j \sim_{i.i.d} \mathcal{N}(0, 10^4)\ j = 0, \ldots, p_\alpha \\ \beta_j \sim_{i.i.d} \mathcal{N}(0, 10^4)\ j = 0, \ldots, p_\beta \end{cases}$$

Une interprétation possible [168] d'un tel modèle est de considérer que la mortalité infantile a une progression en deux étapes : l'IRIS i n'est pas à risque à l'origine (et ne peut donc pas avoir de cas), mais l'influence des variables explicatives du modèle logistique sur π_i peut faire « basculer » cet IRIS parmi les IRIS à risque, alors le nombre de cas observés dans cet IRIS suivra une loi de Poisson (et pourra être influencé par les variables explicatives sur θ_i). L'observation d'un nombre de cas nul dans un IRIS pourrait alors être dû à deux possibilités : soit l'IRIS n'est pas à risque, soit il est à risque mais aucun cas n'a été observé. Les paramètres du modèle logistique s'interpréteraient alors comme des odds-ratios du risque qu'un IRIS devienne à risque et les paramètres du modèle log-linéaire comme des risques relatifs de mortalité dans les IRIS à risque.

11.3 Discussion

Après un examen rapide des possibilités des modèles ZIP, ceux-ci nous sont apparus comme potentiellement utiles dans notre cadre d'étude. Malheureusement, il n'a pas été possible d'appliquer ce modèle en temps imparti pour des raisons techniques (d'implémentation de celui-ci dans le logiciel WinBUGS0 causant des problèmes de convergence des chaînes de Markov) afin d'en étudier plus précisément l'intérêt.

Une limite de ces modèles en l'état actuel de notre réflexion semble cependant être que s'ils peuvent gérer la variabilité extra-poissonnienne de nos données, ils ne semblent pas pouvoir prendre en compte l'aspect spatial des données, qui est pourtant l'un des objectifs de cette thèse et l'une des particularités de l'étude.

Leur application fait néanmoins figure de perspective possible pour étudier les liens entre les différentes variables explicatives et les taux de mortalité infantile de manière différente.

Conclusion et discussion de la troisième partie

Dans cette partie, nous nous sommes concentrés sur l'objectif de la prise en compte de la structure spatiale des données dans des modèles explicatifs d'événements rares. Faute de temps, nous nous sommes principalement concentrés sur le modèle BYM pour atteindre cet objectif, tout en explorant néanmoins les possibilités offertes par les modèles ZIP.

La prise en compte de la structure spatiale des données fait partie des problématiques méthodologiques qui se posent désormais fréquemment dans les contextes d'applications de l'épidémiologie sociale et de l'épidémiologie environnementale. En effet, de plus en plus d'études sont désormais menées en ayant comme échelle des unités géographiques, souvent définies administrativement, sur lesquelles les différentes données sont mesurées. Néanmoins, cette structure spatiale implique souvent une autocorrélation, où des unités voisines ont des caractéristiques proches et des unités éloignées des caractéristiques plus différentes, qu'il convient donc de prendre en compte.

Plusieurs techniques (modèle BYM, modèles additifs généralisés, modèles de balayage spatial) sont désormais utilisées fréquemment dans ces domaines pour répondre à cet objectif mais de nombreuses études continuent néanmoins d'utiliser les techniques d'analyses « traditionnelles » conçues pour des données individuelles (ou sans structure spatiale). Ceci nous a conduit à comparer à plusieurs reprises les résultats obtenus par ces différents modèles et ceux des modèles que nous avons étudiés, que ce soit avant le début de cette thèse (voir article C.1) ou au cours de celle-ci (voir article C.2 et section 10.3).

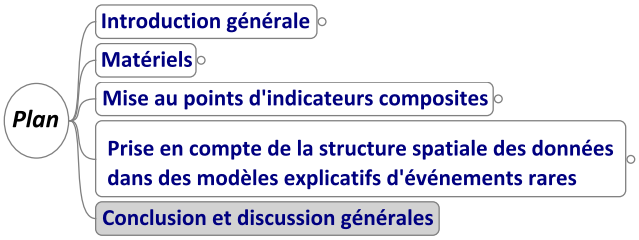
Nous avons pu constater que le modèle BYM permet d'inclure à la fois des termes d'hétérogénéité structurée et non structurée, effectuant ainsi à la fois un lissage global et local des estimations des risques de mortalité infantile. Nous avons vu que ce modèle permettait un meilleur ajustement des données et provoquait de légers changements dans les estimations des effets des variables explicatives par rapport à un modèle de Poisson « classique ». Le modèle BYM demeure cependant plus complexe à mettre en œuvre, notamment à cause du contexte bayésien dans lequel il se place ou du choix des lois *a priori* à effectuer, et plus compliqué à communiquer.

Nous avons également exploré la possibilité d'employer des modèles ZIP pour répondre à l'objectif de cette partie. Malgré leur prise en compte de la variabilité extra-poissonnienne des données, ils ne semblent cependant pas, en l'état de nos réflexions, adaptés à notre objectif et au contexte d'application.

Pour conclure, il nous semble important de poursuivre les réflexions sur le modèle BYM, dont l'intérêt nous paraît certain. La prise en compte des connaissances déjà existantes via la définition de lois *a priori* informatives ; l'utilisation d'autres définitions du voisinage permettant de mieux refléter cette notion ; ou encore l'exploitation de l'ensemble des informations contenues dans les distributions *a posteriori* plutôt qu'uniquement des indicateurs ponctuels et un intervalle crédible nous apparaissent comme d'importantes pistes de réflexions pouvant améliorer encore davantage l'utilité de ce modèle dans un contexte de santé publique, et ce malgré les difficultés d'utilisation qu'il peut occasionner.

Conclusion et discussion générales

Conclusion et discussion générales



Comme nous l'avons développé dans ce manuscrit, cette thèse a pour but d'améliorer les connaissances concernant différentes méthodologies dans un domaine d'étude dans lequel elles ont été peu mises en œuvre, et de déterminer si l'apport de ces techniques statistiques compense les inconvénients d'une mise en place plus complexe par rapport aux techniques « traditionnellement » utilisées.

Plus particulièrement, deux volets principaux composaient cette recherche :

- d'une part la construction d'indicateurs composites capables de synthétiser l'information contenue dans des bases de données importantes ;
- d'autre part la prise en compte de la structure spatiale de ces données au sein de modèles explicatifs d'événements rares.

Pour soutenir nos travaux et illustrer nos réponses à ces objectifs, nous nous sommes inscrit dans un projet visant à explorer les liens entre les inégalités sociales, les expositions environnementales et la mortalité infantile et néonatale. Cette illustration a été réalisée dans les quatre plus grandes agglomérations françaises : Paris et sa Petite Couronne, Marseille, Lyon et Lille choisies pour leurs contrastes et leur taille. Les études ont été faites à l'échelle de l'IRIS, la plus petite échelle géographique d'analyse possible actuellement, afin de réduire le biais écologique et d'avoir une interprétation fine des résultats.

Pour créer ces indicateurs composites, nous avons utilisé des techniques d'analyse de données classiques mais dont l'emploi est encore peu répandu dans ce champ d'application.

Nous avons tout d'abord mis au point une procédure de création d'indices socio-économiques qui permet, partant d'un ensemble de variables socio-économiques, d'effectuer une sélection et une synthèse de celles-ci en se basant sur des ACP successives. L'analyse de données est utilisée ici afin d'avoir une approche dirigée par les données, statistiquement justifiée et réduisant la subjectivité de choix des variables incluses dans l'indice socio-économique

final. Cette approche permet également d'inclure davantage de variables que de nombreux indices existants tout en gardant la possibilité d'interpréter finement les liens entre l'indice et chacune de ces variables. Il s'agit d'un outil pouvant être utilisé aussi bien en recherche épidémiologique que par des acteurs et décideurs locaux. Les indices créés par cette procédure demeurent cependant des indices composites (dont l'interprétation peut être compliquée) à une échelle agrégée (ce qu'il faut prendre en compte si on les combine à des données individuelles).

Nous avons ensuite créé un indice de multi-expositions environnementales qui s'inspire de cette procédure et permet de synthétiser des expositions environnementales différentes de manière « simple » pour déterminer différents types de « fardeau environnemental » pesant sur les IRIS. Le choix s'est porté ici sur l'AFM car nous avons souhaité donner un poids identique à chaque type d'exposition. Là encore, cet outil peut être utilisé aussi bien directement par des acteurs locaux qu'en tant que variable explicative dans des modèles pour l'épidémiologie. On a constaté que cet indice partage les inconvénients précédemment cités des indices composites, et ne permet en outre pas de prendre en compte les effets sanitaires réels de chaque exposition.

Enfin, nous avons exploré la possibilité d'une analyse de données multi-niveaux dans le cas des indicateurs socio-économiques. Bien que cette technique nous permette de construire des indices socio-économiques différenciant les contrastes entre l'échelle communale et l'échelle des IRIS, son interprétation difficile et la différence de but d'utilisation nous a conduit à ne pas développer davantage cet aspect.

Nous avons néanmoins pu montrer que la mise au point d'indicateurs composites à l'aide de techniques d'analyse de données permettait d'exploiter les différentes bases de données mises à disposition des chercheurs d'une manière efficace et utile dans le domaine d'application de la santé publique, et ceci malgré une complexité légèrement supérieure aux méthodes utilisées généralement.

Le second volet de cette thèse a consisté à explorer les méthodes de prise en compte de la structure spatiale des données, en particulier pour l'explication du risque d'événements rares. Ceci nous a conduit principalement à étudier et appliquer le modèle bayésien « BYM ».

L'utilisation de la statistique bayésienne présente ici comme avantages de pouvoir définir de manière simple un modèle complexe permettant de prendre en compte l'aspect spatial des données de l'étude. Cette approche permet également d'inclure les connaissances déjà acquises (des experts locaux ou de la littérature, par exemple) dans le modèle via la distribution *a priori* des paramètres, mais aussi d'avoir d'autres types de résultats que ceux habituellement utilisés grâce à la distribution *a posteriori* obtenue. L'une des limitations majeure de ce modèle est qu'il s'inscrit dans un cadre bayésien qui est méconnu, complexe à appréhender et dont l'application est plus coûteuse en temps. Malgré ces difficultés nécessitant d'importants efforts de communication et de réalisation, le modèle BYM permet de prendre en compte la structure spatiale des données de manière efficace.

Nous avons également exploré la possibilité d'utiliser les modèles ZIP pour répondre à notre problématique, mais ceux-ci ne semblent pas y être adaptés.

Ces deux volets ont été illustrés par de nombreuses applications issues principalement du projet Equit'Area permettant de montrer l'utilité de ces techniques dans un contexte de santé publique.

En termes de perspectives, outre celles présentées dans les différentes discussions (extension du package R, ajout d'autres expositions environnementales, modifications des distributions *a priori* et de la définition du voisinage, etc), les différents éléments de cette thèse ne présentent pas tous les mêmes degrés d'extension possibles. La procédure de création d'indices socio-économiques semble par exemple pouvoir (et a déjà commencé à) être étendue et appliquée à de nombreux autres contextes, bien que des réflexions doivent être menées dans certains cas comme les zones rurales. Le modèle BYM semble également pouvoir être également utilisé et développé davantage malgré sa complexité d'utilisation dans le domaine. *A contrario*, l'indice de multi-expositions semble à l'heure actuelle limité dans ses extensions possibles.

Pour conclure, nous avons au cours de cette thèse été à la frontière entre les mathématiques appliquées et la santé publique pour explorer les deux objectifs principaux cités précédemment.

Tout au long de ce travail, une attention particulière a été apportée au transfert de techniques parfois « classiques » de statistiques dans un domaine d'application où elles étaient moins connues.

Nous avons ainsi pu montrer que malgré certains inconvénients dont souvent une complexité plus importante de mise en œuvre, ces techniques pouvaient apporter beaucoup au champ de la santé publique.

Il semble par conséquent important que ces efforts de transmission se poursuivent afin de permettre une utilisation plus fréquente et appropriée de ces techniques dans les applications.

Bibliographie

[1] Mackenbach, J. P., Cavelaars, A. E., Kunst, A. E. et Groenhof, F. **Socioeconomic inequalities in cardiovascular disease mortality ; an international study**. *European Heart Journal* 21.14, p. 1141–1151 (2000).

[2] Steenland, K., Henley, J. et Thun, M. **All-cause and cause-specific death rates by educational status for two million people in two American Cancer Society cohorts, 1959-1996**. *American Journal of Epidemiology* 156.1, p. 11–21 (2002).

[3] Chaix, B., Rosvall, M. et Merlo, J. **Recent increase of neighborhood socioeconomic effects on ischemic heart disease mortality : a multilevel survival analysis of two large Swedish cohorts**. *American Journal of Epidemiology* 165.1, p. 22–26 (2007).

[4] Tassone, E. C., Waller, L. A. et Casper, M. L. **Small-Area Racial Disparity in Stroke Mortality**. *Epidemiology* 20, p. 234–241 (2009).

[5] Borrell, C., Marí-Dell'olmo, M., Serral, G., Martínez-Beneito, M. et Gotsens, M. **Inequalities in mortality in small areas of eleven Spanish cities (the multicenter MEDEA project)**. *Health & Place* 16.4, p. 703–711 (2010).

[6] Prescott, E, Godtfredsen, N, Vestbo, J et Osler, M. **Social position and mortality from respiratory diseases in males and females**. *The European Respiratory Journal : Official Journal of the European Society for Clinical Respiratory Physiology* 21.5, p. 821–826 (2003).

[7] Ellison-Loschmann, L, Sunyer, J, Plana, E, Pearce, N, Zock, J.-P., Jarvis, D, Janson, C, Antó, J. M. et Kogevinas, M. **Socioeconomic status, asthma and chronic bronchitis in a large community-based study**. *The European Respiratory Journal : Official Journal of the European Society for Clinical Respiratory Physiology* 29.5, p. 897–905 (2007).

[8] McLaren, L. **Socioeconomic Status and Obesity**. *Epidemiologic Reviews* 29.1, p. 29–48 (2007).

[9] Donnelly, D. W. et Gavin, A. **Socio-economic inequalities in cancer incidence - The choice of deprivation measure matters**. *Cancer Epidemiology* (2011).

[10] Meijer, M., Röhl, J., Bloomfield, K. et Grittner, U. **Do neighborhoods affect individual mortality ? A systematic review and meta-analysis of multilevel studies**. *Social Science & Medicine* 74.8, p. 1204–1212 (2012).

[11] Tello, J. E., Jones, J., Bonizzato, P., Mazzi, M., Amaddeo, F. et Tansella, M. **A census-based socio-economic status (SES) index as a tool to examine the relationship between mental health services use and deprivation**. *Social Science & Medicine (1982)* 61.10, p. 2096–2105 (2005).

[12] Curtis, S., Copeland, A., Fagg, J., Congdon, P., Almog, M. et Fitzpatrick, J. **The ecological relationship between deprivation, social isolation and rates of hospital admission for acute psychiatric care : a comparison of London and New York City**. *Health & Place* 12.1, p. 19–37 (2006).

[13] Pattenden, S, Dolk, H et Vrijheid, M. **Inequalities in low birth weight : parental social class, area deprivation, and "lone mother" status**. *Journal of Epidemiology and Community Health* 53.6, p. 355–358 (1999).

[14] Krieger, N, Chen, J. T., Waterman, P. D., Soobader, M.-J., Subramanian, S. V. et Carson, R. **Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning : The Public Health Disparities Geocoding Project (US)**. *Journal of Epidemiology and Community Health* 57.3, p. 186–199 (2003).

[15] Arntzen, A., Samuelsen, S. O., Bakketeig, L. S. et Stoltenberg, C. **Socioeconomic status and risk of infant death. A population-based study of trends in Norway, 1967-1998**. *International Journal of Epidemiology* 33.2, p. 279–288 (2004).

[16] Singh, G. K. et Kogan, M. D. **Persistent socioeconomic disparities in infant, neonatal, and postneonatal mortality rates in the United States, 1969-2001**. *Pediatrics* 119.4, e928–939 (2007).

[17] Marmot, M. **Social determinants of health inequalities**. *Lancet* 365.9464, p. 1099–1104 (2005).

[18] Cambois, E., Laborde, C. et Robine, J.-M. **La « double peine » des ouvriers : plus d'années d'incapacité au sein d'une vie plus courte**. *Population & Sociétés* 441 (2008).

[19] Blanpain, N. **L'espérance de vie s'accroît, les inégalités sociales face à la mort demeurent**. *Insee Première* 1372 (2011).

[20] Marmot, M. **Achieving health equity : from root causes to fair outcomes**. *The Lancet* 370.9593, p. 1153–1163 (2007).

[21] WHO. *Global Health Observatory Data Repository*. 2013. URL : http://apps.who.int/gho/data/node.main.

[22] *The World Bank - World DataBank*. URL : http://databank.worldbank.org.

[23] WHO. *World Health Statistics 2007*. Geneva, Switzerland : World Health Organization, 2007.

[24] *CIA - The World Factbook*. URL : https://www.cia.gov/library/publications/the-world-factbook/index.html.

[25] United Nations. *The Millennium Development Goals Report 2013*. New York : United Nations Pubns, 2013.

[26] *Health, United States, 2012 : With Special Feature on Emergency Care*. Rapp. tech. Hyattsville, MD : National Center for Health Statistics, 2013.

[27] Geoff Bascand. *New Zealand Period Life Tables : 2010–12*. Rapp. tech. Statistics New Zealand, 2013.

[28] Lang, T. et Leclerc, A. **Les inégalités sociales de santé en France : portrait épidémiologique**. In : *Réduire les inégalités sociales en santé*. Santé en action. Saint-Denis : Inpes, 2010, p. 62–72.

[29] *ObEpi-Roche, enquête épidémiologique de référence sur l'évolution de l'obésité et du surpoids en France*. 2012. url : http://www.roche.fr/home/recherche/domaines_therapeutiques/cardio_metabolisme/enquete_nationale_obepi_2012.html.

[30] Beck, F., Guignard, R., Richard, J.-B., Wilquin, J.-L. et Peretti-Watel, P. **Augmentation récente du tabagisme en France : principaux résultats du Baromètre santé, France, 2010**. *Bulletin Epidemiologique Hebdomadaire* 20-21, p. 230–231 (2011).

[31] Guignon, N. **La santé des adolescents scolarisés en classe de troisième en 2003-2004 - Premiers résultats**. *DREES Etudes et résultats* 573 (2007).

[32] Euzenat, D. **L'exposition des salariés aux maladies professionnelles en 2007**. *Dares Analyses* 056 (2010).

[33] Barceló, M. A., Saez, M. et Saurina, C. **Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain**. *The Science of the Total Environment* 407.21, p. 5501–5523 (2009).

[34] Deguen, S., Lalloué, B., Bard, D., Havard, S., Arveiler, D. et Zmirou-Navier, D. **A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex : a Bayesian modeling approach**. *Epidemiology (Cambridge, Mass.)* 21.4, p. 459–466 (2010).

[35] Torjesen, I. **Health professions pledge action against socioeconomic factors responsible for health inequalities**. *BMJ (Clinical research ed.)* 346, f1814 (2013).

[36] WHO. *Millennium Development Goals*. url : http://www.who.int/topics/millennium_development_goals/en/.

[37] Sixty-fifth World Health Assembly. *WHA65.8 - Outcome of the World Conference on Social Determinants of Health*. 2012.

[38] European Parliament and European Council. **Decision 1350/2007/EC of the European Parliament and of the Council of 23 October 2007 establishing a second programme of Community action in the field of health (2008-13)**. *Official Journal of the European Union* 301.3 (2007).

[39] Commission of the European Communities. *Solidarity in Health : Reducing Health Inequalities in the EU COM(2009) 567 final*. Rapp. tech. Commission of the European Communities, 2009.

[40] Fifth Ministerial Conference on Environment and Health. *Parma Declaration on Environment and Health*. Rapp. tech. EUR/55934/5.1 Rev. 2. Parma, Italy : WHO Regional Office for Europe, 2010.

[41] **LOI n° 2004-806 du 9 août 2004 relative à la politique de santé publique**. *Journal Officiel de la République Française* 185, p. 14277 (2004).

[42] Ministère du Travail, de l'Emploi et de la Santé. *Programme national nutrition santé 2011-2015*. Rapp. tech. 2011.

[43] Ministère de l'Enseignement Supérieur et de la Recherche et Ministère de la Santé et des Sports. *Plan cancer 2009-2013*. Rapp. tech. 2009.

[44] Basset, B. *Agences régionales de santé - Les inégalités sociales de santé*. Saint-Denis : INPES, 2009.

[45] Potvin, L., Moquet, M.-J. et Jones, C. M. *Réduire les inégalités sociales en santé*. Santé en action. Saint-Denis : Inpes, 2010.

[46] *La Santé de l'Homme - Inégalités sociales de santé : connaissances et modalités d'intervention*. La Santé de l'Homme 414. INPES, 2011.

[47] *Les inégalités sociales de santé : sortir de la fatalité*. Rapp. tech. Haut Conseil de la Santé Publique, 2009.

[48] *Indicateurs de suivi de l'évolution des inégalités sociales de santé dans les systèmes d'information en santé*. Rapp. tech. Haut Conseil de la Santé Publique, 2013.

[49] *Avis relatif à la réduction des inégalités sociales et territoriales de santé Place des Ateliers santé ville*. Rapp. tech. Haut Conseil de la Santé Publique, 2013.

[50] *Avis relatif aux indicateurs de suivi de l'évolution des inégalités sociales de santé dans le domaine du cancer*. Rapp. tech. Haut Conseil de la Santé Publique, 2013.

[51] Murray, C. J., Salomon, J. A. et Mathers, C. **A critical examination of summary measures of population health.** *Bulletin of the World Health Organization* 78.8, p. 981–994 (2000).

[52] Parrish, R. G. **Measuring Population Health Outcomes**. *Preventing Chronic Disease* 7.4 (2010).

[53] Reidpath, D et Allotey, P. **Infant mortality rate as an indicator of population health**. *Journal of Epidemiology and Community Health* 57.5, p. 344–346 (2003).

[54] Mathews, T. J. et MacDorman, M. F. **Infant mortality statistics from the 2007 period linked birth/infant death data set**. *National vital statistics reports : from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 59.6, p. 1–30 (2011).

[55] EURO-PERISTAT Project with SCPE and EUROCAT. *European Perinatal Health Report. The health and care of pregnant women and babies in Europe in 2010*. Rapp. tech. 2013.

[56] Guildea, Z. E., Fone, D. L., Dunstan, F. D., Sibert, J. R. et Cartlidge, P. H. **Social deprivation and the causes of stillbirth and infant mortality**. *Archives of Disease in Childhood* 84.4, p. 307–310 (2001).

[57] Niel, X. *Les facteurs explicatifs de la mortalité infantile en France et leur évolution récente*. Document de travail F1106. INSEE, 2011.

[58] EURO-PERISTAT Project, with SCPE, EUROCAT, EURONEOSTAT. *European Perinatal Health Report*. Rapp. tech. 2008.

[59] INSEE. *Institut national de la statistique et des études économiques*. URL : http://www.insee.fr.

[60] Wilkinson, R. G. et Marmot, M. G. *Social determinants of health. The solid facts (second edition)*. Copenhagen : World Health Organization, Regional Office for Europe, 2003.

[61] Luo, Z.-C., Wilkins, R. et Kramer, M. S. **Effect of neighbourhood income and maternal education on birth outcomes : a population-based study**. *CMAJ : Canadian Medical Association Journal = Journal De l'Association Medicale Canadienne* 174.10, p. 1415–1420 (2006).

[62] Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T. et Sidorchuk, A. **Type 2 diabetes incidence and socio-economic position : a systematic review and meta-analysis**. *International Journal of Epidemiology* 40.3, p. 804–818 (2011).

[63] Hiscock, R., Bauld, L., Amos, A., Fidler, J. A. et Munafò, M. **Socioeconomic status and smoking : a review**. *Annals of the New York Academy of Sciences* 1248.1, 107–123 (2012).

[64] Evans, G. W. et Kantrowitz, E. **Socioeconomic Status and Health : The Potential Role of Environmental Risk Exposure**. *Annual Review of Public Health* 23.1, p. 303–331 (2002).

[65] O'Neill, M. S., Jerrett, M., Kawachi, I., Levy, J. I., Cohen, A. J., Gouveia, N., Wilkinson, P., Fletcher, T., Cifuentes, L. et Schwartz, J. **Health, wealth, and air pollution : advancing theory and methods**. *Environmental health perspectives* 111.16, p. 1861–1870 (2003).

[66] Brown, P. **Race, Class, and Environmental Health : A Review and Systematization of the Literature**. *Environmental Research* 69.1, p. 15–30 (1995).

[67] Finkelstein, M. M., Jerrett, M., DeLuca, P., Finkelstein, N., Verma, D. K., Chapman, K. et Sears, M. R. **Relation between income, air pollution and mortality : a cohort study**. *CMAJ : Canadian Medical Association Journal = Journal De l'Association Medicale Canadienne* 169.5, p. 397–402 (2003).

[68] Briggs, D., Abellan, J. J. et Fecht, D. **Environmental inequity in England : small area associations between socio-economic status and environmental pollution**. *Social Science & Medicine (1982)* 67.10, p. 1612–1629 (2008).

[69] Braubach, M. et Fairburn, J. **Social inequities in environmental risks associated with housing and residential location–a review of evidence**. *European Journal of Public Health* 20.1, p. 36–42 (2010).

[70] Viel, J.-F., Hägi, M., Upegui, E. et Laurian, L. **Environmental justice in a French industrial region : Are polluting industrial facilities equally distributed ?** *Health & Place* 17, p. 257–262 (2011).

[71] Forastiere, F., Stafoggia, M., Tasco, C., Picciotto, S., Agabiti, N., Cesaroni, G. et Perucci, C. A. **Socioeconomic status, particulate air pollution, and daily mortality : Differential exposure or differential susceptibility**. *American Journal of Industrial Medicine* 50.3, p. 208–216 (2007).

[72] Havard, S., Deguen, S., Zmirou-Navier, D., Schillinger, C. et Bard, D. **Traffic-related air pollution and socioeconomic status : a spatial autocorrelation study to assess environmental equity on a small-area scale**. *Epidemiology (Cambridge, Mass.)* 20.2, p. 223–230 (2009).

[73] Ferrer, R. L. **Pursuing Equity : Contact With Primary Care and Specialist Clinicians by Demographics, Insurance, and Health Status**. *The Annals of Family Medicine* 5.6, p. 492–502 (2007).

[74] O'Neill, M. S., Zanobetti, A. et Schwartz, J. **Modifiers of the Temperature and Mortality Association in Seven US Cities**. *American Journal of Epidemiology* 157.12, p. 1074–1082 (2003).

[75] Yi, O., Kim, H. et Ha, E. **Does area level socioeconomic status modify the effects of PM10 on preterm delivery ?** *Environmental Research* 110.1, p. 55–61 (2010).

[76] *Projet de recherche Equit'Area*. 2013. URL : http://www.equitarea.org.

[77] European Council. **Council Decision 2005/370/EC of 17 February 2005 on the conclusion, on behalf of the European Community, of the Convention on access to information, public participation in decision-making and access to justice in environmental matters**. *Official Journal of the European Communities* (2005).

[78] *Réseau Quetelet - Réseau français des centres de données pour les sciences sociales*. URL : http://www.reseau-quetelet.cnrs.fr/spip/.

[79] Ministère de l'Ecologie, du Développement Durable et de l'Energie. *Eider - Base de données régionales et départementales sur l'environnement, l'énergie, le transport, le logement et la construction*. URL : http://www.stats.environnement.developpement-durable.gouv.fr/Eider/.

[80] *Plateforme française d'ouverture des données publiques (Open Data)*. URL : http://www.data.gouv.fr/.

[81] Anthopolos, R., James, S. A., Gelfand, A. E. et Miranda, M. L. **A spatial measure of neighborhood level racial isolation applied to low birthweight, preterm birth, and birthweight in North Carolina**. *Spatial and Spatio-temporal Epidemiology* 2.4, p. 235–246 (2011).

[82] Eibner, C. et Sturm, R. **US-based indices of area-level deprivation : results from HealthCare for Communities**. *Social Science & Medicine (1982)* 62.2, p. 348–359 (2006).

[83] Marí-Dell'Olmo, M., Martínez-Beneito, M. A., Borrell, C., Zurriaga, O., Nolasco, A. et Domínguez-Berjón, M. F. **Bayesian factor analysis to calculate a deprivation index and its uncertainty**. *Epidemiology (Cambridge, Mass.)* 22.3, p. 356–364 (2011).

[84] O'Campo, P, Xue, X, Wang, M. C. et Caughy, M. **Neighborhood risk factors for low birthweight in Baltimore : a multilevel analysis**. *American Journal of Public Health* 87.7, p. 1113–1118 (1997).

[85] INSEE. *Les Ilots Regroupés pour des Indicateurs Statistiques 2008*. URL : http://www.insee.fr/fr/ppp/bases-de-donnees/donnees-detaillees/duicq/accueil.asp?page=doc/iris2008.htm.

[86] Krieger, N. **Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence : Does the Choice of Area-based Measure and Geographic Level Matter ? : The Public Health Disparities Geocoding Project**. *American Journal of Epidemiology* 156.5, p. 471–482 (2002).

[87] Greenland, S. et Morgenstern, H. **Ecological bias, confounding, and effect modification**. *International journal of epidemiology* 18.1, 269–274 (1989).

[88] Sénat - Service des études juridiques. *Étude de législation comparée n° 184 - avril 2008 - Les enfants nés sans vie.* 2008. url : http://www.senat.fr/lc/lc184/lc1840.html.

[89] Bell, M. L., Ebisu, K. et Belanger, K. **Ambient air pollution and low birth weight in Connecticut and Massachusetts**. *Environmental health perspectives* 115.7, p. 1118–1124 (2007).

[90] Medeiros, A. P. P. de, Gouveia, N., Machado, R. P. P., Souza, M. R. de, Alencar, G. P., Novaes, H. M. D. et Almeida, M. F. de. **Traffic-related air pollution and perinatal mortality : a case-control study**. *Environmental health perspectives* 117.1, p. 127–132 (2009).

[91] Inserm. *CépiDC - Centre d'épidémiologie sur les causes médicales de décès.* url : http://www.cepidc.inserm.fr/site4/.

[92] INSEE. *Le recensement - Pourquoi la méthode a-t-elle changé en 2004 ?* url : http://www.insee.fr/fr/publics/default.asp?page=communication/recensement/particuliers/changement.htm.

[93] **LOI n° 2002-276 du 27 février 2002 relative à la démocratie de proximité**. *Journal Officiel de la République Française*, p. 3808 (2002).

[94] INSEE. *Le recensement - L'organisation du recensement.* url : http://www.insee.fr/fr/publics/default.asp?page=communication/recensement/particuliers/organisation.htm.

[95] Jarman, B. **Identification of underprivileged areas.** *British Medical Journal (Clinical research ed.)* 286.6379, p. 1705–1709 (1983).

[96] Townsend, P., Phillimore, P. et Beattie, A. *Health and deprivation : inequality and the North.* London Routledge, 1988.

[97] Carstairs, V. et Morris, R. **Deprivation : explaining differences in mortality between Scotland and England and Wales.** *BMJ : British Medical Journal* 299.6704, p. 886–889 (1989).

[98] Carstairs, V. et Morris, R. **Deprivation and mortality : an alternative to social class ?** *Journal of Public Health* 11.3, p. 210 –219 (1989).

[99] Morris, R et Carstairs, V. **Which deprivation ? A comparison of selected deprivation indexes**. *Journal of Public Health Medicine* 13.4, p. 318–326 (1991).

[100] Carstairs, V. **Deprivation indices : their interpretation and use in relation to health**. *Journal of Epidemiology and Community Health* 49 Suppl 2, S3–8 (1995).

[101] Salmond, C, Crampton, P et Sutton, F. **NZDep91 : A New Zealand index of deprivation**. *Australian and New Zealand Journal of Public Health* 22.7, p. 835–837 (1998).

[102] Messer, L. C., Laraia, B. A., Kaufman, J. S., Eyster, J., Holzman, C., Culhane, J., Elo, I., Burke, J. G. et O'Campo, P. **The development of a standardized neighborhood deprivation index**. *Journal of Urban Health : Bulletin of the New York Academy of Medicine* 83.6, p. 1041–1062 (2006).

[103] Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W. et Davey Smith, G. **Indicators of socioeconomic position (part 2)**. *Journal of Epidemiology and Community Health* 60.2, p. 95–101 (2006).

[104] Pampalon, R, Hamel, D, Gamache, P et Raymond, G. **A deprivation index for health planning in Canada**. *Chronic Diseases in Canada* 29.4, p. 178–191 (2009).

[105] Rey, G., Jougla, E., Fouillet, A. et Hémon, D. **Ecological association between a deprivation index and mortality in France over the period 1997 - 2001 : variations with spatial scale, degree of urbanicity, age, gender and cause of death**. *BMC Public Health* 9, p. 33 (2009).

[106] INSEE. *Unité de consommation*. URL : http://www.insee.fr/fr/methodes/default.asp?page=definitions/unite-consommation.htm.

[107] Moran, P. A. P. **Notes on Continuous Stochastic Phenomena**. *Biometrika* 37.1/2, p. 17 (1950).

[108] *Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide*. Rapp. tech. Bonn, Germany : WHO Regional Office for Europe, 2003.

[109] WHO. *Air Quality Guidelines Global Update 2005*. Rapp. tech. Bonn, Germany : WHO Regional Office for Europe, 2005.

[110] *AirParif*. URL : http://www.airparif.asso.fr/.

[111] *Air PACA*. URL : http://www.atmopaca.org/.

[112] *Air Rhône-Alpes*. URL : http://www.air-rhonealpes.fr.

[113] *Atmo Nord-Pas de Calais*. URL : http://www.atmo-npdc.fr/home.htm.

[114] *AirParif - Méthodes de surveillance - Le réseau de mesure*. URL : http://www.airparif.asso.fr/methodes-surveillance/reseau-mesure.

[115] McHugh, C., Carruthers, D. et Edmunds, H. **ADMS-Urban : an air quality management system for traffic, domestic and industrial pollution**. *International Journal of Environment and Pollution* 8.3-4, p. 666–674 (1997).

[116] Soulhac, L., Salizzoni, P., Cierco, F.-X. et Perkins, R. **The model SIRANE for atmospheric urban pollutant dispersion ; part I, presentation of the model**. *Atmospheric Environment* 45.39, p. 7379–7395 (2011).

[117] Soulhac, L., Salizzoni, P., Mejean, P., Didier, D. et Rios, I. **The model SIRANE for atmospheric urban pollutant dispersion ; PART II, validation of the model on a real case study**. *Atmospheric Environment* 49, p. 320–337 (2012).

[118] TARGETING, KTT, TUV. *Street 5.2 : Logiciel d'évaluation simple de la pollution atmosphérique provoquée par la circulation automobile*. Rapp. tech. 2005.

[119] WHO. *Guidelines for community noise. 1999*. Rapp. tech. World Health Organization, : Geneva, 2008.

[120] European Parliament and European Council. **Directive 2002/49/EC of the European Parliament and of the Council of 25 June 2002 relating to the assessment and management of environmental noise**. *Official Journal of the European Communities* (2002).

[121] *Centre Scientifique et Technique du Bâtiment*. URL : http://www.cstb.fr/.

[122] Brender, J. D., Maantay, J. A. et Chakraborty, J. **Residential Proximity to Environmental Hazards and Adverse Health Outcomes**. *American Journal of Public Health* 101.S1, S37–S52 (2011).

[123] Downey, L. et Van Willigen, M. **Environmental Stressors : The Mental Health Impacts of Living Near Industrial Activity**. *Journal of health and social behavior* 46.3, p. 289–305 (2005).

[124] *E-PRTR - The European Pollutant Release and Transfer Register*. URL : http://prtr.ec.europa.eu/.

[125] Beckerman, B., Jerrett, M., Brook, J. R., Verma, D. K., Arain, M. A. et Finkelstein, M. M. **Correlation of nitrogen dioxide with other traffic pollutants near a major expressway**. *Atmospheric Environment* 42.2, p. 275–290 (2008).

[126] *Institut national de l'information géographique et forestière*. URL : http://www.ign.fr/.

[127] Mitchell, R. et Popham, F. **Greenspace, urbanity and health : relationships in England**. *Journal of Epidemiology and Community Health* 61.8, p. 681–683 (2007).

[128] Maas, J., Verheij, R. A., Vries, S. d., Spreeuwenberg, P., Schellevis, F. G. et Groenewegen, P. P. **Morbidity is related to a green living environment**. *Journal of Epidemiology and Community Health* 63.12, p. 967–973 (2009).

[129] Ministère de l'Ecologie, du Développement Durable et de l'Energie. *Occupation des sols - CORINE Land Cover*. URL : http://www.statistiques.developpement-durable.gouv.fr/donnees-ligne/liste/1825.html.

[130] Stephansson, O, Dickman, P. W., Johansson, A. L. et Cnattingius, S. **The influence of socioeconomic status on stillbirth risk in Sweden**. *International Journal of Epidemiology* 30.6, p. 1296–1301 (2001).

[131] Braveman, P. A., Cubbin, C., Egerter, S., Chideya, S., Marchi, K. S., Metzler, M. et Posner, S. **Socioeconomic status in health research : one size does not fit all**. *JAMA : The Journal of the American Medical Association* 294.22, p. 2879–2888 (2005).

[132] Gordon, D. **Census based deprivation indices : their weighting and validation**. *Journal of Epidemiology and Community Health* 49 Suppl 2, S39–44 (1995).

[133] Krieger, N, Williams, D. R. et Moss, N. E. **Measuring social class in US public health research : concepts, methodologies, and guidelines**. *Annual Review of Public Health* 18, p. 341–378 (1997).

[134] Fukuda, Y., Nakamura, K. et Takano, T. **Higher mortality in areas of lower socioeconomic position measured by a single index of deprivation in Japan**. *Public Health* 121.3, p. 163–173 (2007).

[135] Bell, N., Schuurman, N. et Hayes, M. V. **Using GIS-based methods of multicriteria analysis to construct socio-economic deprivation indices**. *International Journal of Health Geographics* 6, p. 17 (2007).

[136] Pornet, C., Delpierre, C., Dejardin, O., Grosclaude, P., Launay, L., Guittet, L., Lang, T. et Launoy, G. **Construction of an adaptable European transnational ecological deprivation index : the French version**. *Journal of Epidemiology & Community Health* (2012).

[137] Havard, S., Deguen, S., Bodin, J., Louis, K., Laurent, O. et Bard, D. **A small-area index of socioeconomic deprivation to capture health inequalities in France**. *Social Science & Medicine (1982)* 67.12, p. 2007–2016 (2008).

[138] Singh, G. K. **Area deprivation and widening inequalities in US mortality, 1969-1998**. *American Journal of Public Health* 93.7, p. 1137–1143 (2003).

[139] Cohen, J. **A Coefficient of Agreement for Nominal Scales**. *Educational and Psychological Measurement* 20.1, p. 37–46 (1960).

[140] Rand, W. M. **Objective Criteria for the Evaluation of Clustering Methods**. *Journal of the American Statistical Association* 66.336, p. 846 (1971).

[141] R Development Core Team. **R : A language and environment for statistical computing**. *R Foundation for Statistical Computing Vienna Austria* (2011).

[142] Lê, S., Josse, J. et Husson, F. **FactoMineR : An R package for multivariate analysis**. *Journal of statistical software* 25.1, 1–18 (2008).

[143] *The Comprehensive R Archive Network*. URL : http://cran.r-project.org/.

[144] *BruitParif - Projet SURVOL*. URL : http://www.bruitparif.fr/projet-survol.

[145] Passchier-Vermeer, W et Passchier, W. F. **Noise exposure and public health.** *Environmental Health Perspectives* 108.Suppl 1, p. 123–131 (2000).

[146] Callahan, M. A. et Sexton, K. **If Cumulative Risk Assessment Is the Answer, What Is the Question ?** *Environmental Health Perspectives* 115.5, p. 799–806 (2007).

[147] Pearce, J. R., Richardson, E. A., Mitchell, R. J. et Shortt, N. K. **Environmental justice and health : A study of multiple environmental deprivation and geographical inequalities in health in New Zealand**. *Social Science & Medicine* 73, p. 410–420 (2011).

[148] Sexton, K. **Cumulative Risk Assessment : An Overview of Methodological Approaches for Evaluating Combined Health Effects from Exposure to Multiple Environmental Stressors**. *International Journal of Environmental Research and Public Health* 9.12, p. 370–390 (2012).

[149] U.S. EPA. *Framework for Cumulative Risk Assessment*. Rapp. tech. EPA/630/P02/001F. Washington DC : U.S. Environmental Protection Agency, 2003.

[150] California Environmental Justice Action Plan. *Environmental Justice Action Plan*. Rapp. tech. Cal-EPA, 2004.

[151] Commission of the European Communities. *The European Environment & Health Action Plan 2004-2010 COM(2004) 416 final*. text/html ; charset=UTF-8. Commission of the European Communities, 2004.

[152] National Research Council (U.S.) *Science and decisions advancing risk assessment*. Washington, D.C. : National Academies Press, 2009.

[153] Su, J. G., Morello-Frosch, R., Jesdale, B. M., Kyle, A. D., Shamasunder, B. et Jerrett, M. **An index for assessing demographic inequalities in cumulative environmental hazards with application to Los Angeles, California**. *Environmental Science & Technology* 43.20, p. 7626–7634 (2009).

[154] Richardson, E. A., Mitchell, R., Shortt, N. K., Pearce, J. et Dawson, T. P. **Developing summary measures of health-related multiple physical environmental deprivation for epidemiological research [Abstract only]**. *Environment and Planning A* 42.7, p. 1650–1668 (2010).

[155] Yorita Christensen, K. L. et White, P. **A methodological approach to assessing the health impact of environmental chemical mixtures : PCBs and hypertension in the National Health and Nutrition Examination Survey**. *International journal of environmental research and public health* 8.11, p. 4220–4237 (2011).

[156] Huang, G. et London, J. K. **Cumulative Environmental Vulnerability and Environmental Justice in California's San Joaquin Valley**. *International Journal of Environmental Research and Public Health* 9.5, p. 1593–1608 (2012).

[157] Escofier, B. et Pagès, J. **Multiple factor analysis (AFMULT package)**. *Computational Statistics & Data Analysis* 18.1, p. 121–140 (1994).

[158] Zeka, A., Melly, S. J. et Schwartz, J. **The effects of socioeconomic status and indices of physical environment on reduced birth weight and preterm births in Eastern Massachusetts**. *Environmental Health : A Global Access Science Source* 7, p. 60 (2008).

[159] Lebart, L., Piron, M. et Morineau, A. *Statistique exploratoire multidimensionnelle : visualisations et inférences en fouille[s] de données*. Paris : Dunod, 2006.

[160] Christensen, R., Johnson, W., Branscum, A. et Hanson, T. E. *Bayesian ideas and data analysis : an introduction for scientists and statisticians*. Texts in Statistical Science. Boca Raton, FL : CRC Press, 2011.

[161] Robert, C. P. *Le choix bayésien Principes et pratique*. Paris : Springer-Verlag France, 2006.

[162] Ntzoufras, I. *Bayesian modeling using WinBUGS*. Hoboken, N.J. : Wiley, 2009.

[163] Lawson, A. B., Browne, W. J. et Vidal Rodeiro, C. L. *Disease mapping with WinBUGS & MLwiN*. Statistics in Practice. Chichester : Wiley, 2003.

[164] Elliott, P, Wakefield, J., Best, N. et Briggs, D. *Spatial epidemiology : methods and applications*. Oxford ; New York : Oxford University Press, 2001.

[165] Besag, J., York, J. et Mollié, A. **Bayesian image restoration, with two applications in spatial statistics**. *Annals of the Institute of Statistical Mathematics* 43.1, p. 1–20 (1991).

[166] Lunn, D. J., Thomas, A., Best, N. et Spiegelhalter, D. **WinBUGS-a Bayesian modelling framework : concepts, structure, and extensibility**. *Statistics and computing* 10.4, 325–337 (2000).

[167] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. et Van Der Linde, A. **Bayesian measures of model complexity and fit**. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 64.4, 583–639 (2002).

[168] Cheung, Y. B. **Zero-inflated models for regression analysis of count data : a study of growth and development**. *Statistics in Medicine* 21.10, 1461–1469 (2002).

[169] Bayes, M. et Price, M. **An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS**. *Philosophical Transactions (1683-1775)*, 370–418 (1763).

[170] Laplace, P.-S. d. **Mémoire sur la Probabilité des Causes par les Événements**. *Mémoires de l'Académie royale des Sciences de Paris (Savants étrangers)* VI, p. 621–656 (1774).

[171] Metropolis, N. et Ulam, S. **The Monte Carlo Method**. *Journal of the American Statistical Association* 44.247, p. 335 (1949).

[172] Von Neumann, J. **Various techniques used in connection with random digits**. *J Resources of the National Bureau of Standards-Applied Math Series* 12.36-38, p. 1 (1951).

[173] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. et Teller, E. **Equation of State Calculations by Fast Computing Machines**. *The Journal of Chemical Physics* 21.6, p. 1087–1092 (1953).

[174] Hastings, W. K. **Monte Carlo sampling methods using Markov chains and their applications**. *Biometrika* 57.1, p. 97 –109 (1970).

[175] Chib, S. et Greenberg, E. **Understanding the Metropolis-Hastings Algorithm**. *The American Statistician* 49.4, p. 327–335 (1995).

[176] Geman, S. et Geman, D. **Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6.6, p. 721–741 (1984).

[177] Gelman, A. et Rubin, D. B. **Inference from Iterative Simulation Using Multiple Sequences**. *Statistical Science* 7.4, p. 457–472 (1992).

Annexes

Annexe A

Rappels méthodologiques

A.1 Analyse de données

A.1.1 Analyse en composantes principales

L'analyse en composantes principales est une technique d'analyse de données qui vise à décrire (mettre en évidence les similarités et dissimilarités entre les individus statistiques et les corrélations entre variables), synthétiser (déterminer un faible nombre de nouvelles variables, combinaisons linéaires non corrélées de variance maximum des variables d'origine) et visualiser les informations contenues dans un tableau de données.

Soit X une table de données de n lignes représentant les individus statistiques et p colonnes représentant les variables (on considère que celles-ci sont centrées). On note x_i^j la valeur de la variable j pour l'individu i. On note s^j l'écart type de la j^e variable. On note D la matrice diagonale des poids p_i des individus (généralement, on a $p_i = \frac{1}{n}\ \forall i$). Afin d'atteindre les buts précédemment cités, l'analyse est réalisée en trois étapes.

Dans la première étape, le but est de construire une représentation graphique visualisable des individus qui soit telle que ceux ayant des valeurs proches soient représentés par des points proches et que les individus ayant des valeurs éloignées soient représentés par des points éloignés. Soit A_i le point de coordonnées $(x_i^1, \ldots, x_i^p)$ dans $\mathbb{R}^p$, représentant l'individu i.

On représente les différences entre individus par une distance euclidienne d entre les points les représentant. Pour cela, il est nécessaire de choisir une métrique M dans $\mathbb{R}^p$ permettant de définir d. Généralement, la métrique M choisie est la matrice diagonale de l'inverse des variances, ce qui revient à prendre pour distance entre les points A_i et $A_{i'}$: $d(A_i, A_{i'}) = \sqrt{\sum_{j=1}^{p} \frac{1}{(s^j)^2}(x_i^j - x_{i'}^j)^2}$. Cette distance a notamment pour avantage de ne pas dépendre de l'unité de mesure des variables et de les placer sur un même ordre de grandeur.

Une fois la représentation des individus et la distance entre les points déterminées, le but de l'ACP est de pouvoir visualiser ces points. Comme le nombre de variables est généralement important, il est nécessaire de projeter le nuage de points sur un sous-espace visualisable, choisi de façon à ce que la projection soit la plus fidèle possible au nuage de points initial.

On va donc chercher un sous-espace de projection F_r de dimension r tel que les distances entre les projections P_i des points A_i sur F_r soient les plus proches des distances entre les

points A_i, c'est-à-dire tel que $\sum_{i,i'=1}^{n} p_i p_{i'} \|\overrightarrow{P_i P_{i'}}\|^2$ soit maximale. Si l'on choisit le sous-espace F_r tel qu'il contienne le barycentre G des points A_i, ceci est équivalent à rendre $\sum_{i=1}^{n} p_i \|\overrightarrow{GP_i}\|^2$maximale. Cela revient à chercher le sous-espace affine F_r de dimension r qui ajuste le mieux le nuage de points (A_i, p_i) au sens des moindres carrés.

La détermination d'une base orthonormée (par rapport à M) $(\overrightarrow{u_1}, \ldots, \overrightarrow{u_r})$ de F_r se fait de manière itérative en constatant que :

- $\sum_{i=1}^{n} p_i \|\overrightarrow{GP_i}\|^2 = \sum_{k=1}^{r} \underline{u}_k' MX'DXM\underline{u}_k$ par l'application des théorèmes des trois perpendiculaires et de Pythagore, et car la base $(\overrightarrow{u_1}, \ldots, \overrightarrow{u_r})$ est M-orthonormée.
- Il est possible à partir d'un sous-espace F_{r-1} vérifiant le critère de construire un sous-espace F_r le vérifiant également en ajoutant à F_{r-1} une droite orthogonale définie par $\overrightarrow{u_r}$ et telle que $\underline{u}_r' MX'DXM\underline{u}_r$ soit maximale sous la contrainte $\underline{u}_r' M\underline{u}_r = 1$
- $\underline{u}_k$ est vecteur propre de $X'DXM$ associé à sa k^e plus grande valeur propre λ_k.

On obtient ainsi pas à pas une base de F_r et on appelle l'axe $(G, \overrightarrow{u_k})$ le k^e axe principal. À partir de cette analyse, il est déjà possible d'obtenir de nombreuses informations sur la qualité de représentation des différents points (ou du nuage entier) sur le sous-espace F_r et sur les axes principaux. Il est également possible de représenter graphiquement la projection du nuage de points sur ce sous-espace.

La deuxième étape permet de donner une interprétation statistique. Le k^e axe principal est ainsi la représentation d'une combinaison linéaire des variables d'origine appelée k^e facteur principal. Plus précisément, le k^e facteur principal est la combinaison linéaire des variables d'origine (de coefficients $\underline{a}_k = M\underline{u}_k$) telle que sa valeur pour l'individu i soit l'abscisse de la projection du point A_i sur le k^e axe principal.

On peut interpréter statistiquement le k^e facteur principal comme la combinaison linéaire (des variables d'origine) de variance maximale sous la contrainte d'être non corrélée aux facteurs précédents. La qualité globale de représentation sur le k^e axe principal, c'est-à-dire le ratio $\frac{\lambda_k}{\lambda_1 + \ldots + \lambda_p}$, peut également être interprétée comme le pourcentage de variance expliquée par le k^e facteur.

La troisième et dernière étape consiste en une étude dans l'espace des variables $\mathbb{R}^n$. Les variables sont ainsi représentées par les points B_j de coordonnées $(x_1^j, \ldots, x_n^j)$. On analyse ces points de manière similaire aux points A_i dans la première étape : au lieu d'étudier le tableau X avec la métrique M et la matrice de poids D, on va étudier le tableau X' avec la métrique D et la matrice de poids M.

Cette étape permet d'étudier les liens entre variables et de déterminer un sous-espace de projection des variables, utilisé pour construire la représentation graphique nommée « cercle de corrélations ». Cette représentation permet de visualiser les corrélations entre les variables, entre les variables et les facteurs principaux, mais aussi la qualité de représentation des variables.

L'interprétation de l'ACP se déroule ainsi de manière à la fois graphique, en étudiant les différentes projections du nuage de points des individus sur les axes principaux (et les plans qu'ils constituent) et les projections des variables (cercle de corrélations), et numérique, via l'étude des corrélations entre variables et facteurs, des coordonnées des individus sur les axes, de leur contribution à ceux-ci, etc.

A.1.2 Analyse factorielle multiple

Soit $X = (X^1|\cdots|X^q)$ un tableau de données composé d'observations sur les mêmes n individus statistiques de q groupes de variables, et soit m_k le nombre de variables du sous-ensemble X^k.

La première étape de l'AFM est, pour chaque sous-ensemble X^k $k = 1, 2, \ldots, q$, de réaliser une ACP si le groupe est composé de variables quantitatives ou une AFCM si le groupe est composé de variables qualitatives. Soit M^k la métrique utilisée pour l'analyse du groupe k, c'est-à-dire la matrice utilisée pour calculer la distance entre les points représentant les individus, et soit λ_1^k la plus grande valeur propre de l'analyse sur le groupe k.

La deuxième étape de l'AFM est alors de réaliser une ACP sur le tableau X avec la métrique M, où M est la matrice $\begin{pmatrix} \frac{M^1}{\lambda_1^1} & & \\ & \ddots & \\ & & \frac{M^q}{\lambda_1^q} \end{pmatrix}$. Ceci permet de donner en un certain sens le même poids à chaque groupe de variable, même s'ils ont des tailles très différentes. L'interprétation de l'AFM est ensuite similaire à celle d'une ACP.

A.1.3 Classification ascendante hiérarchique

La classification ascendante hiérarchique est une technique de classification non supervisée, fréquemment utilisée après une analyse factorielle.

Soit I un ensemble de n éléments ($I = 1, 2, \ldots, n$) et P l'ensemble des poids $p_1, \ldots, p_n$ affectés à ces éléments. On note A_i le point de $\mathbb{R}^p$ représentant l'individu i. Soit d une distance entre points définie par la métrique M.

On souhaite trouver une partition de I en r classes (notées $I_1, \ldots, I_r$) qui maximise l'inertie inter-classes, c'est-à-dire $\sum_{k=1}^{r} P_k d^2(G_k, G)$, où $P_k = \sum_{i \in I_k} p_i$, G_k est le barycentre des points appartenant à la classe k et G est le barycentre de l'ensemble des points. On notera que ceci est équivalent à minimiser l'inertie intra-classes $\sum_{k=1}^{r} \sum_{i \in I_k} p_i d^2(A_i, G_k)$. Ce critère permet de créer des classes homogènes dans leur composition et hétérogènes entre elles.

Cependant, le nombre de partitions possibles de n éléments en k classes augmente extrêmement vite avec n et rend la recherche d'une solution directe irréalisable, ce qui conduit à utiliser une méthode approchée. Pour cela, l'algorithme de classification ascendante hiérarchique est utilisé avec une distance Δ entre classes (basée sur d) particulière : la distance de Ward. La distance de Ward entre les classes I_j et I_k est définie comme $\Delta(I_j, I_k) = \frac{P_j P_k}{P_j + P_k} d^2(G_j, G_k)$.

L'algorithme de CAH est le suivant :

- Étape 1 : À partir de la partition contenant tous les singletons $P_0 = \{\{1\}, \{2\}, \ldots, \{n\}\}$, on calcule la distance Δ entre toutes les paires de singletons. Les classes $\{l\}$ et $\{m\}$ ayant la distance Δ minimale sont réunies, ce qui conduit à obtenir la partition de $(n-1)$ éléments $P_1 = \{\{l, m\}, \{1\}, \ldots, \{n\}\}$.
  ...

- Étape r : À partir de la partition P_{r-1} contenant $(n-(r-1))$ éléments, la distance est calculée entre les éléments de la partition (en pratique, il suffit de calculer la distance entre la nouvelle classe constituée à l'étape précédente et les autres éléments). Les classes de distance Δ minimale sont regroupées et la partition de $(n-r)$ éléments, P_r est créée. Par définition de la distance de Ward, le regroupement de ces deux classes est celui qui minimise la perte d'inertie inter-classes parmi tous les autres regroupements possibles à cette étape.
  ...
- Étape $n-1$: la partition créée est $P_{n-1} = \{I\}$

Les résultats des regroupements effectués par CAH sont généralement représentés par un dendrogramme qui indique pour chaque étape le regroupement qui a été fait et la perte d'inertie inter-classes qui y est liée. La première étape de l'interprétation est de choisir où couper l'arbre, c'est-à-dire de choisir quelle partition sera utilisée. Ensuite, il est possible d'interpréter les classes à l'aide, par exemple, des statistiques descriptives des différentes variables au sein de chaque classe ou encore par l'étude des éléments composant chacune.

A.2 Généralités sur la statistique bayésienne

L'approche statistique classique (ou « fréquentiste ») est depuis environ un siècle fortement prédominante dans l'enseignement et l'utilisation des statistiques. L'approche bayésienne est longtemps restée peu utilisée pour des raisons (entre autres) historiques et techniques, en particulier à cause des difficultés analytiques et de calcul qu'elle peut occasionner. Cependant, l'essor de l'informatique lui a permis de se développer et d'être désormais régulièrement employée dans de nombreux domaines [160, 161], dont les analyses spatiales. Il n'entre pas dans le cadre de cette section de comparer ces deux approches et leurs pertinences respectives, ce point faisant l'objet de nombreux débats (relevant parfois autant de la philosophie que des mathématiques), mais de présenter quelques généralités sur l'approche bayésienne et son application.

A.2.1 L'approche bayésienne

L'approche bayésienne tire son nom du révérend Thomas Bayes, qui a vécu durant la première moitié du XVIIIe siècle et que l'on connait en particulier grâce au théorème portant son nom (voir section A.2.2). Cependant, bien que ce théorème soit également crucial dans le raisonnement bayésien, c'est surtout par la manière qu'a eu Bayes de considérer que le paramètre p d'une loi de probabilité (dans le cas d'une loi binomiale) pouvait lui-même suivre une loi de probabilité qu'il a ouvert la voie à l'approche bayésienne au sens actuel du terme.

En effet, il est habituel en statistiques de considérer des modèles de probabilité pour les données contenant un certain nombre de paramètres,ceux-ci ne pouvant jamais être connus avec une certitude absolue (à moins de travailler sur la population entière plutôt que sur un échantillon). En statistique classique, on considère ces paramètres comme des valeurs fixes inconnues que l'on va chercher à estimer (avec une certaine précision).

L'approche bayésienne va considérer que, en plus du modèle statistique sur les données, les paramètres inconnus de ce modèle suivent eux-mêmes des lois de probabilité destinées à modéliser l'incertitude sur ceux-ci. Pour cela, on utilisera des informations connues *a priori* (basées sur l'information scientifique disponible, l'expertise, la littérature, d'autres données, etc.) obtenues de manière indépendante des données. C'est ce qui conduit l'approche bayésienne à être « basée sur le postulat que toute incertitude doit être modélisée en utilisant des probabilités et que les inférences statistiques doivent être des conclusions logiques basées sur les lois de probabilité » (traduction de Christensen *et al.* [160]). Ceci est par ailleurs lié à l'interprétation bayésienne des probabilités comme une traduction numérique d'un degré de connaissance et pas forcément comme des fréquences d'occurrence d'événements répétés un très grand nombre de fois.

A.2.2 Le théorème de Bayes

Avant de poursuivre, rappelons le théorème de Bayes, établi d'abord par Thomas Bayes [169](présenté à titre posthume en 1763) puis indépendamment redécouvert et étendu en 1774 par Pierre-Simon de Laplace [170]. Sous sa forme probablement la plus connue, on peut le présenter de la manière suivante :

Théorème de Bayes

Soit A et B deux événements aléatoires de probabilité non nulle. La probabilité conditionnelle de A sachant B est définie par $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

On a :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A})}$$

Ce théorème existe également dans une version qui nous intéresse davantage, utilisant des densités de probabilité :

Théorème de Bayes (densités)

Soit deux vecteurs aléatoires $\underline{\theta} = (\theta_1, \ldots, \theta_r)'$ et $\underline{Y} = (Y_1, \ldots, Y_n)'$.

On note :

- $f_{\underline{\theta},\underline{Y}}(\underline{u}, \underline{v}) = f_{\underline{\theta},\underline{Y}}(u_1, \ldots, u_r, v_1, \ldots, v_n)$ la densité jointe du vecteur aléatoire $(\underline{\theta}', \underline{Y}')'$
- $f_{\underline{\theta}}(\underline{u})$ (respectivement $f_{\underline{Y}}(\underline{v})$) la densité marginale de $\underline{\theta}$ (resp. $\underline{Y}$) :

$$f_{\underline{\theta}}(\underline{u}) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{\underline{\theta},\underline{Y}}(u_1, \ldots, u_r, v_1, \ldots, v_n) dv_1 \ldots dv_n$$

- $f_{(\underline{\theta}|\underline{Y})}(\underline{u}|\underline{v})$ (resp. $f_{(\underline{Y}|\underline{\theta})}(\underline{v}|\underline{u})$) la densité conditionnelle de $\underline{\theta}$ sachant $\underline{Y}$ (resp. de $\underline{Y}$ sachant $\underline{\theta}$) :

$$f_{(\underline{\theta}|\underline{Y})}(\underline{u}|\underline{v}) = \frac{f_{\underline{\theta},\underline{Y}}(\underline{u}, \underline{v})}{f_{\underline{Y}}(\underline{v})}$$

On a alors :

$$f_{(\underline{\theta}|\underline{Y})}(\underline{u}|\underline{v}) = \frac{f_{(\underline{Y}|\underline{\theta})}(\underline{v}|\underline{u}) f_{\underline{\theta}}(\underline{u})}{\int f_{(\underline{Y}|\underline{\theta})}(\underline{v}|\underline{u}) f_{\underline{\theta}}(\underline{u}) d\underline{u}}$$

Afin d'alléger les notations, on notera :

$$f(\underline{\theta}|\underline{Y}) = \frac{f(\underline{Y}|\underline{\theta}) f(\underline{\theta})}{\int f(\underline{Y}|\underline{\theta}) f(\underline{\theta}) d\underline{\theta}}$$

ou même

$$f(\underline{\theta}|\underline{Y}) \propto f(\underline{Y}|\underline{\theta}) f(\underline{\theta})$$

le dénominateur n'étant qu'une constante de normalisation.

En pratique, on considère que l'on a des données observées $\underline{y}$ issues du vecteur aléatoire $\underline{Y}$ suivant une loi d'échantillonnage de paramètre $\underline{\theta}$ et de densité $f(\underline{Y}|\underline{\theta})$. On a ainsi la vraisemblance $f(\underline{y}|\underline{\theta})$. On donne au paramètre $\underline{\theta}$ une loi *a priori* de densité $f(\underline{\theta})$. À partir de la vraisemblance et de la loi *a priori*, on peut alors obtenir en utilisant le théorème de Bayes la loi *a posteriori* $f(\underline{\theta}|\underline{y})$ de $\underline{\theta}$.

Il s'agit donc en réalité de la formalisation probabiliste d'un raisonnement naturel : partant d'une connaissance (ou d'une opinion) *a priori*, on confronte celle-ci aux données disponibles afin de la mettre à jour et d'obtenir alors une connaissance *a posteriori* prenant en compte les données.

A.2.3 Lois *a priori*

Le choix de la distribution *a priori* de $\underline{\theta}$ fait partie des points importants de l'approche bayésienne (et c'est aussi l'une de ses principales différences pratiques avec l'approche classique). Cette distribution doit en effet exprimer la connaissance que l'on a déjà sur le paramètre, indépendamment des données. On remarquera cependant que suivant la forme fonctionnelle de la vraisemblance et de la loi *a priori*, l'obtention de la loi *a posteriori* sous une forme explicite de densité de loi connue peut être particulièrement compliquée, voire impossible. C'est principalement pour cette raison que l'approche bayésienne est longtemps restée confinée à des cas particuliers où il était possible d'obtenir cette loi, avant que le développement de l'informatique ne généralise l'utilisation des méthodes d'approximation numérique (dont il sera discuté ci-dessous) pour permettre d'employer des distributions *a priori* plus complexes.

Le choix de loi *a priori* a longtemps été cantonné pour des raisons pratiques à l'utilisation de lois particulières : les lois conjuguées. Si la distribution *a posteriori* est de la même famille de distributions que la distribution *a priori*, on dira ainsi qu'elles sont conjuguées, et la loi *a priori* sera dite loi *a priori* conjuguée par rapport à la vraisemblance. On peut ainsi montrer que si la vraisemblance est une loi $\mathcal{P}(\theta)$ et que la loi *a priori* de θ est une loi $\Gamma(a, b)$, alors la loi *a posteriori* sera la loi $\Gamma(a + \sum_{i=1}^{n} y_i, b + n)$.

Choisir une loi *a priori* conjuguée à la vraisemblance permet une importante simplification des calculs, puisque l'on connaît alors directement la forme de la loi *a posteriori* et que

l'on peut calculer facilement ses paramètres. Ceci explique leur utilisation presque systématique avant l'essor de l'informatique. Cependant, le choix de la loi *a priori* relevait alors davantage de considérations pratiques que d'une réelle modélisation des connaissances.

Le développement des techniques de simulation numérique permettant de simuler des échantillons de lois de probabilité et d'estimer des intégrales compliquées a permis de se libérer de cette contrainte. Il est désormais possible de baser le choix des lois *a priori* soit sur des connaissances passées ou l'avis d'un expert (en pratiquant alors ce que les anglophones nomment *elicitation*), soit sur des distributions *a priori* « de référence », appelées parfois également « non informatives », dans le cas où il n'y a pas de connaissances *a priori* disponibles.

Le processus d'*elicitation* étant un procédé long et complexe (faisant l'objet de nombreuses illustrations détaillées dans Christensen *et al.* [160]) qui implique de traduire les connaissances d'un expert en lois de probabilité, l'utilisation de lois *a priori* « de référence » est souvent préférée. Ces lois sont choisies pour fournir une base commune d'évaluation des données sans modéliser d'information *a priori* spécifique, et parce qu'elles affectent très peu la loi *a posteriori* (d'où le nom également utilisé de loi *a priori* « non informative ») par rapport aux données. Ainsi, on utilise fréquemment comme lois *a priori* de référence :

- pour les paramètres de moyennes : des lois uniformes, des lois plates impropres (dont l'intégrale de la densité n'est pas égale à 1) ou encore des lois normales centrées de variance très importante (10^3 ou 10^4)
- pour les termes de variance, des lois gamma inverse ayant des paramètres petits (10^{-2} ou 10^{-3}).

Cependant, on peut constater que les lois *a priori* dépendent généralement elles-mêmes de paramètres et que ces paramètres peuvent avoir une incertitude. C'est ce qui conduit aux modèles hiérarchiques.

A.2.4 Modèles hiérarchiques

Une fois que l'on admet que le paramètre $\underline{\theta}$ dispose d'une loi de probabilité *a priori* $f(\underline{\theta})$, la construction de modèles hiérarchiques se fait de manière naturelle. En effet, $f(\underline{\theta})$ va généralement posséder certains paramètres $\underline{\phi}$ sur lesquels il existe une incertitude, il est donc également possible de modéliser celle-ci par une loi *a priori* $f(\underline{\phi})$ (nommée *hyperprior*). Cette nouvelle loi *a priori* peut elle-même avoir d'autres paramètres, fixes ou sur lesquels on a une incertitude modélisée par une loi *a priori*, etc.

On peut alors définir un modèle par « étages » successifs (d'où le terme de modèle hiérarchique, ou parfois de modèles multi-niveaux) permettant d'exprimer de manière simple des modèles menant à des lois *a posteriori* complexes. Dans le cas d'un modèle à deux niveaux, on peut ainsi avoir comme formulation :

- Le modèle sur les données : $\underline{Y}|\underline{\theta},\underline{\phi} \sim f(\underline{Y}|\underline{\theta},\underline{\phi})$ (on effectue ici un léger abus de notation pour signifier que $\underline{Y}|\underline{\theta},\underline{\phi}$ suit une loi de probabilité de densité $f(\underline{Y}|\underline{\theta},\underline{\phi})$). On considèrera généralement que $\underline{Y}$ est indépendante de $\underline{\theta}|\underline{\phi}$ et donc que $f(\underline{Y}|\underline{\theta}) = f(\underline{Y}|\underline{\theta},\underline{\phi})$
- Le modèle *a priori* sur les paramètres : $\underline{\theta}|\underline{\phi} \sim f(\underline{\theta}|\underline{\phi})$

- L'hyperprior $\underline{\phi} \sim f(\underline{\phi})$ (cette distribution ayant elle-même des paramètres que l'on considérera ici fixés).

On notera que cette formulation n'est qu'une facilité d'écriture d'un modèle bayésien « classique », puisque l'on peut ramener la formulation précédente :

- Soit à un modèle avec pour densité d'échantillonnage $f(\underline{Y}|\underline{\theta}) = \int f(\underline{Y}|\underline{\theta}, \underline{\phi}) f(\underline{\theta}|\underline{\phi}) d\underline{\theta}$ et densité *a priori* $f(\underline{\phi})$
- Soit à un modèle avec pour densité d'échantillonnage $f(\underline{Y}|\underline{\theta}, \underline{\phi})$ et densité *a priori* $f(\underline{\theta}) = \int f(\underline{\theta}|\underline{\phi}) f(\underline{\phi}) d\underline{\phi}$

Ces deux autres formulations ne changent pas la structure du modèle, mais la formulation hiérarchique est plus simple et pratique.

De cette formulation on obtient facilement la densité *a posteriori* par le théorème de Bayes :

$$f(\underline{\theta}, \underline{\phi}|\underline{y}) \propto f(\underline{y}|\underline{\theta}) f(\underline{\theta}|\underline{\phi}) f(\underline{\phi})$$

Si l'on souhaite plus particulièrement faire une inférence sur le paramètre θ_i, on utilisera alors la densité marginale *a posteriori* :

$$f(\theta_i|\underline{y}) = \int_{\underline{\phi}} \int_{\theta_1} \cdots \int_{\theta_{i-1}} \int_{\theta_{i+1}} \cdots \int_{\theta_r} f(\underline{\theta}, \underline{\phi}|\underline{y}) d\underline{\phi} d\theta_1 \ldots d\theta_{i-1} d\theta_{i+1} \ldots d\theta_r$$

Cette formulation (dont on retrouve des équivalents en statistique classique avec les modèles à variables latentes, par exemple) a divers avantages [161] :

- Elle permet de refléter des connaissances objectives sur la modélisation du phénomène observé. C'est ainsi que les modèles hiérarchiques peuvent être utilisés dans le cadre d'analyses « multi-niveaux » impliquant des données sur différentes échelles géographiques, ou encore dans le cadre de méta-analyses.
- Elle permet également de modéliser l'information *a priori* en plusieurs niveaux de distributions conditionnelles. Ceci peut permettre de refléter une séparation entre un *a priori* « structurel » et un *a priori* plus « subjectif ».
- Elle permet d'exprimer de manière simple des modèles complexes et également de simplifier les calculs numériques effectués.

A.2.5 Méthodes de Monte-Carlo par chaînes de Markov

Ainsi que nous l'avons évoqué précédemment, c'est grâce à l'apport des méthodes de simulation numérique qu'il est désormais possible d'obtenir des résultats sur des lois *a posteriori* qu'il est souvent difficile d'exprimer en pratique sous une forme explicite connue, en permettant d'évaluer les différentes intégrales impliquées dans le processus d'inférence bayésienne. Ces intégrations sont d'autant plus complexes à effectuer que le nombre de paramètres est élevé.

La technique couramment employée pour atteindre ce but est celle de Monte-Carlo par chaînes de Markov (MCMC, pour *Markov chain Monte Carlo*), dont voici quelques rappels [161]. Pour alléger l'écriture, on n'utilisera pas de barre dans la notation des vecteurs dans cette section, bien que le paramètre θ puisse parfaitement être multidimensionnel.

Rappels sur la méthode de Monte-Carlo

La méthode de Monte-Carlo, introduite par Metropolis, Ulan et von Neumann [171, 172] permet d'approximer la valeur d'une intégrale $I = \int g(x)dx$ en faisant appel à une fonction de densité f que l'on sait simuler. En effet, on peut alors obtenir une estimation de I en simulant un échantillon $x^{(1)}, \ldots, x^{(T)}$ de loi de probabilité de densité f et en considérant alors la moyenne $\hat{I} = \frac{1}{T}\sum_{t=1}^{T} \frac{g(x^{(t)})}{f(x^{(t)})}$ (cette somme convergeant presque sûrement vers $\mathbb{E}\left[\frac{g(X)}{f(X)}\right] = I$ lorsque T tend vers l'infini).

On voit bien ainsi que si g désigne toute fonction d'intérêt du paramètre θ, on peut facilement obtenir une estimation de son espérance *a posteriori* $\mathbb{E}[g(\theta)|y] = \int g(\theta)f(\theta|y)d\theta$ en simulant un échantillon $\theta^{(1)}, \ldots, \theta^{(T)}$ suivant la distribution *a posteriori* $f(\theta|y)$ puis en calculant $\hat{I} = \frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)})$. Notons que l'on peut également utiliser les simulations pour estimer la distribution *a posteriori* de $g(\theta)$.

La difficulté est cependant de simuler un échantillon suivant la loi *a posteriori* $f(\theta|y)$, et c'est ici qu'interviennent les chaînes de Markov.

Rappels sur la convergence d'une chaîne de Markov[160]

Une chaîne de Markov est un processus stochastique $\{\theta^{(1)}, \ldots, \theta^{(T)}\}$ possédant la propriété de Markov, c'est-à-dire tel que la distribution de θ au temps $t+1$ (que l'on note $\theta^{(t+1)}$) ne dépend que du temps précédent. Autrement dit, on a pour tout ensemble A :

$$\mathbb{P}(\theta^{(t+1)} \in A|\theta^{(t)}, \ldots, \theta^{(1)}) = \mathbb{P}(\theta^{(t+1)} \in A|\theta^{(t)})\ \forall t$$

Si l'on note $f_{t|.}(\theta^{(t)}|\theta^{(t-1)}, \ldots, \theta^{(1)})$ la densité conditionnelle et $f_t(\theta^{(t)})$ la densité marginale de $\theta^{(t)}$, la propriété de Markov revient donc à dire que $f_{t+1|.}(\theta^{(t+1)}|\theta^{(t)}, \ldots, \theta^{(1)}) = f_{t+1|t}(\theta^{(t+1)}|\theta^{(t)})\ \forall t$. Par conséquent on obtient que

$$\begin{aligned}\mathbb{P}(\theta^{(t+1)} \in A) &= \int_A f_{t+1}(\theta)d\theta \\ &= \int_A \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f_{t+1|t}(\theta^{(t+1)}|\theta^{(t)}) \ldots f_{2|1}(\theta^{(2)}|\theta^{(1)}) \\ &\qquad f_1(\theta^{(1)})d\theta^{(1)} \ldots d\theta^{(t+1)}\end{aligned}$$

Dans la suite, on supposera que la chaîne de Markov est homogène, c'est-à-dire que le mécanisme de transition ne change pas avec le temps, donc que la densité $f_{t+1|t}(\theta^{(t+1)}|\theta^{(t)})$ ne dépend pas de t. On la notera alors simplement $f(\theta^{(t+1)}|\theta^{(t)})$

Par ailleurs, on dit qu'une loi π (de densité $\pi(\theta)$) est stationnaire pour la chaîne de Markov si on a $\pi(\theta) = \int f(\theta|\theta^\star)\pi(\theta^\star)d\theta^\star$. Autrement dit, pour tout t, si $\theta^{(t)} \sim \pi$ alors $\theta^{(t+1)} \sim \pi$ puisque

$$f_{t+1}(\theta)d\theta = \int f(\theta^{(t+1)}|\theta^{(t)})f_t(\theta^{(t)})d\theta^{(t)}$$

Enfin, on peut montrer que si une chaîne de Markov est :

- irréductible (tout ensemble d'états de probabilité non nulle pourra être atteint avec une probabilité non nulle),
- récurrente positive (la probabilité qu'un ensemble soit revisité par la chaîne une infinité de fois est égale à 1)
- apériodique (le retour à un ensemble d'états peut se faire sans que ce soit en suivant une période précise - par exemple tous les deux temps),

alors il existe une unique distribution stationnaire π vers laquelle converge presque sûrement la distribution de $\underline{\theta}^{(t)}$ lorsque t tend vers l'infini, quelle que soit la distribution initiale de $\underline{\theta}^{(1)}$.

On comprend donc que, si l'on est capable de construire une chaîne de Markov ayant une loi de transition $f(\theta^{(t+1)}|\theta^{(t)})$ simple à simuler et telle que sa loi stationnaire soit la loi *a posteriori* $f(\theta|y)$, après un temps suffisant, la chaîne de Markov aura atteint sa loi stationnaire et l'on pourra alors obtenir un échantillon de cette dernière et répondre au problème.

Différents algorithmes existent afin d'implémenter ceci. Parmi les plus courants, on trouve l'algorithme de Metropolis-Hasting et l'échantillonneur de Gibbs.

L'algorithme de Metropolis-Hastings a été proposé par Metropolis *et al.* en 1953 [173] puis étendu par Hastings en 1970 [174]. Il s'agit d'un algorithme très général pour définir une chaîne de Markov satisfaisant aux conditions ci-dessus.

On souhaite donc générer un échantillon de taille T suivant la loi cible $f(\theta|y)$. On suppose que l'on dispose d'une distribution « de proposition » $h(\theta'|\theta)$ permettant de générer des candidats (pour la prochaine valeur de l'échantillon). Si l'on note $\theta^{(t)}$ la valeur générée à l'étape t, l'algorithme de Metropolis-Hastings est alors le suivant :

- On définit une valeur initiale $\theta^{(1)}$, soit en la fixant, soit par un tirage aléatoire suivant une distribution $f_1(\theta^{(1)})$
- A la t^e itération, on dispose donc d'un échantillon $\{\theta^{(1)}, \ldots, \theta^{(t)}\}$ et :
  - On génère un candidat $\theta^\star$ à partir de la loi de proposition $h(\theta^\star|\theta^{(t)})$
  - On définit
$$\alpha = \alpha(\theta^\star, \theta^{(t)}) = \min\left\{1, \frac{f(\theta^\star|y)h(\theta^{(t)}|\theta^\star)}{f(\theta^{(t)}|y)h(\theta^\star|\theta^{(t)})}\right\}$$
  - On pose alors
$$\theta^{(t+1)} = \begin{cases} \theta^\star & \text{avec probabilité } \alpha \\ \theta^{(t)} & \text{avec probabilité } 1-\alpha \end{cases}$$

Notons que par définition de $\alpha(\theta^\star, \theta^{(t)})$ il n'est nécessaire de connaître $f(\theta|y)$ qu'à une constante près, ce qui est précisément le cas lorsque l'on détermine la loi *a posteriori* à partir du théorème de Bayes $f(\theta|y) \propto f(y|\theta)f(\theta)$. On montre (une démonstration peut être trouvée dans Chib et Greenberg [175], ou dans Christensen *et al.* [160] dans le cas discret) que cet algorithme converge vers la loi stationnaire $f(\theta|y)$ quelle que soit la distribution de proposition, bien qu'en pratique cette dernière influe sur la vitesse de convergence.

L'échantillonneur de Gibbs, qui a notamment été utilisé pour la première fois par Geman et Geman [176], peut être considéré comme un cas particulier de l'algorithme de Metropolis-Hastings utilisant comme densité de proposition la distribution conditionnelle *a posteriori* $f(\theta_j|\theta_{-j}, y)$ où $\theta_{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)'$. Avec celle-ci, il n'y a plus de rejet de la proposition et par conséquent les valeurs sont mises à jour à chaque itération. L'algorithme est alors le suivant :

- On définit une valeur initiale $\theta^{(1)}$, soit en la fixant, soit par un tirage aléatoire suivant une distribution $f_1(\theta^{(1)})$
- A la t^e itération, on dispose donc d'un échantillon $\{\theta^{(1)}, \ldots, \theta^{(t)}\}$ et :
  - On pose $\theta^\star = \theta^{(t)}$
  - Pour $j = 1, \ldots, p$ on met à jour $\theta_j^\star \sim f(\theta_j^\star|\theta_{-j}^\star, y)$. En pratique, on procède par étapes successives :

$$\begin{aligned}
\theta_1^\star &\quad \text{à partir de } f(\theta_1^\star|\theta_2^{(t)}, \ldots, \theta_p^{(t)}, y) \\
&\quad \vdots \\
\theta_j^\star &\quad \text{à partir de } f(\theta_j^\star|\theta_1^\star, \ldots, \theta_{j-1}^\star, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, y) \\
&\quad \vdots \\
\theta_p^\star &\quad \text{à partir de } f(\theta_p^\star|\theta_1^\star, \ldots, \theta_{p-1}^\star, y))
\end{aligned}$$

  - On pose alors $\theta^{(t+1)} = \theta^\star$

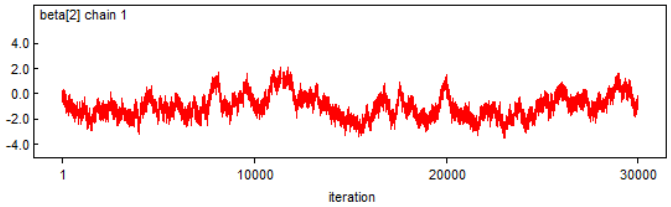
On remarque qu'à l'inverse de l'algorithme de Metropolis-Hastings qui simule à chaque itération l'ensemble du vecteur θ suivant une loi multidimensionnelle, l'algorithme de l'échantillonneur de Gibbs ne simule qu'une composante à la fois à l'aide d'une loi unidimensionnelle (généralement facile à simuler). C'est cette méthode qui est utilisée dans le logiciel WinBUGS [166] que nous avons employé pour effectuer les analyses bayésiennes.

Pour obtenir un échantillon suivant la loi *a posteriori* on va donc, partant de valeurs initiales choisies ou tirées aléatoirement, effectuer de nombreuses itérations de l'un de ces algorithmes. On séparera ensuite les valeurs obtenues en deux : une période de rodage pendant laquelle la chaîne de Markov n'a pas encore convergé et qui ne sera pas conservée, puis une période après la convergence durant laquelle chaque itération de l'algorithme nous fournira un nouveau tirage suivant la loi *a posteriori* que l'on simule.

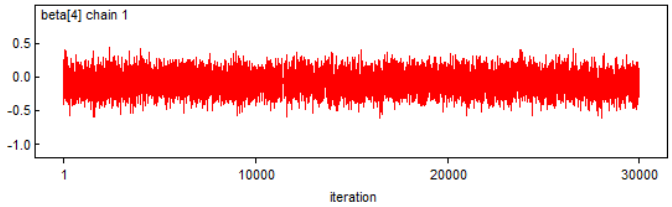
A.2.6 Diagnostics de convergence

Si la convergence est assurée de manière théorique, en pratique rien ne permet de déterminer au bout de combien d'itérations de l'algorithme elle sera atteinte. Pour évaluer si une chaîne a convergé vers sa loi stationnaire, différents diagnostics existent.

On peut par exemple examiner le diagramme des valeurs du paramètre (ou de chacune des composantes du paramètre lorsqu'il est multidimensionnel) en fonction du nombre d'itérations. Suivant la forme de celui-ci, on peut avoir une première information sur la convergence (voir figure A.1). Dans le cas où la convergence est atteinte après la période de rodage, on ne devrait voir aucun motif ni aucune tendance dans le diagramme, la chaîne de Markov se contentant « d'explorer » aléatoirement la distribution *a posteriori*.



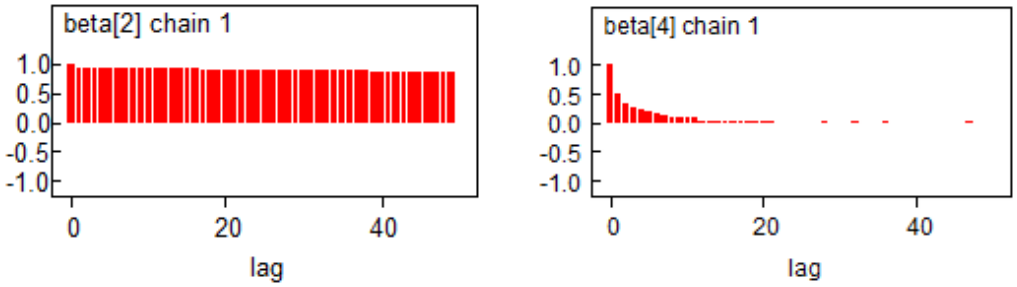
(a) Chaîne n'ayant pas convergé après 30 000 itérations (la chaîne ne s'est pas encore stabilisée)



(b) Chaîne ayant convergé rapidement (la chaîne explore la distribution *a posteriori* aléatoirement)

FIGURE A.1 – Diagrammes des valeurs de deux chaînes de Markov en fonction du nombre d'itérations.

Un second indicateur est la fonction d'autocorrélation qui donne en fonction de j la corrélation entre $\theta^{(t)}$ et $\theta^{(t+j)}$, que l'on représente souvent sous la forme d'un autocorrélogramme (figure A.2). Si la chaîne a convergé, on s'attend à ce que les valeurs soient simulées de manière quasiment non corrélée et par conséquent à ce que l'autocorrélation décroisse très vite avec l'augmentation de j. Un autocorrélogramme où ce n'est pas le cas peut donc indiquer que la chaîne n'a pas encore convergé.



(a) Autocorrélogramme « anormal » indiquant l'absence de convergence

(b) Autocorrélagramme « normal » d'une chaîne ayant convergé

FIGURE A.2 – Autocorrélogramme de deux chaînes de Markov.

Enfin, différents tests de diagnostic existent également, et parmi eux l'un des plus fréquemment utilisé est le diagnostic de Gelman-Rubin. Proposé par Gelman et Rubin en 1992 [177], ce diagnostic se base sur la simulation de plusieurs chaînes de Markov parallèlement partant de différentes valeurs initiales (choisies de manière à avoir une plus grande variance que la distribution *a posteriori*) puis sur une analyse similaire à l'analyse de variance.

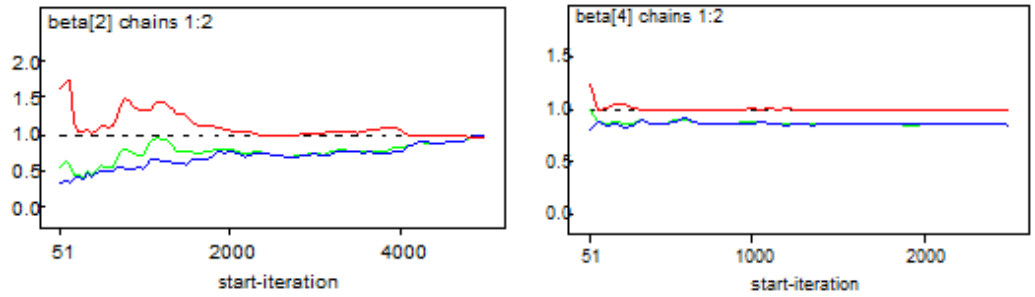
On suppose que l'on dispose de $M \geq 2$ chaînes de Markov chacune de longueur T (la période de rodage ayant déjà été effectuée). On note θ_m^t la valeur au temps t de la chaîne m ; $\bar{\theta}_m^{\cdot} = \frac{1}{T}\sum_{t=1}^{T}\theta_m^t$ la moyenne empirique de θ pour la chaîne m ; $\bar{\theta}_{\cdot}^{\cdot} = \frac{1}{M}\sum_{m=1}^{M}\bar{\theta}_m^{\cdot}$ la moyenne empirique de θ sur l'ensemble des chaînes ; et $s_m^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\theta_m^t - \bar{\theta}_m^{\cdot})^2$ désigne la variance empirique de θ sur la chaîne m.

Gelman et Rubin proposent d'estimer la variance *a posteriori* de deux manières différentes :

- À l'aide de la moyenne des variances empiriques de chaque chaîne, c'est-à-dire la variance intra-chaînes $W = \frac{1}{M}\sum_{m=1}^{M} s_m^2$.
  Cette estimation aura tendance à sous-estimer la variance *a posteriori* tant que les chaînes n'auront pas convergé car les chaînes prises individuellement n'auront pas encore pu parcourir l'ensemble de la loi *a posteriori*.
- Par une combinaison linéaire de la variance intra-chaîne et de la variance inter-chaînes :
  $$\hat{V} = (1 - \frac{1}{T})W + \frac{M+1}{M} \cdot \frac{1}{T}B$$
  où $B = \frac{T}{M-1}\sum_{m=1}^{M}(\bar{\theta}_m^{\cdot} - \bar{\theta}_{\cdot}^{\cdot})^2$ est (T fois) la variance inter-chaîne.
  Cette estimation aura quant à elle tendance à surestimer la variance *a posteriori* tant que la convergence ne sera pas atteinte puisque l'on a pris soin de choisir les valeurs initiales pour cela.

Dans le cas où toutes les chaînes ont convergé (vers la même loi stationnaire : la loi *a posteriori*) on s'attend alors à ce que $\hat{V}$ et W soient proches et stables, et donc que le rapport $\hat{R} = \frac{\hat{V}}{W}$ tende vers 1.

Bien que Gelman et Rubin proposent un test basé sur une approximation de la loi *a posteriori* par une loi de Student, dans la pratique ce diagnostic est souvent réalisé de manière graphique. Ainsi, le logiciel WinBUGS montre l'évolution de $\hat{R}$, de $\hat{V}$ et de W en fonction des itérations, ce qui permet de voir si on a bien $\hat{R} \to 1$ et si $\hat{V}$ et W se rejoignent et se stabilisent. Si ce n'est pas le cas, alors la chaîne n'a pas encore convergé et des itérations supplémentaires sont nécessaires.



(a) Chaîne n'ayant pas encore convergé (b) Chaîne ayant convergé rapidement

FIGURE A.3 – Diagnostic de Gelman-Rubin de deux chaînes de Markov. En rouge $\hat{R}$, en vert $\hat{V}$ et en bleu W

L'ensemble de ces diagnostics s'utilise généralement simultanément afin d'avoir un ensemble d'indicateurs validant le fait que la convergence a bien été atteinte.

Annexe B

Articles principaux de la thèse

Lalloué et al. International Journal for Equity in Health 2013, 12:21
http://www.equityhealthj.com/content/12/1/21

INTERNATIONAL JOURNAL FOR EQUITY IN HEALTH

RESEARCH Open Access

A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis

Benoît Lalloué[1,2,3*], Jean-Marie Monnez[3], Cindy Padilla[1,2], Wahida Kihal[1], Nolwenn Le Meur[1,4], Denis Zmirou-Navier[1,2,5] and Séverine Deguen[1,2]

Abstract

Introduction: In order to study social health inequalities, contextual (or ecologic) data may constitute an appropriate alternative to individual socioeconomic characteristics. Indices can be used to summarize the multiple dimensions of the neighborhood socioeconomic status. This work proposes a statistical procedure to create a neighborhood socioeconomic index.

Methods: The study setting is composed of three French urban areas. Socioeconomic data at the census block scale come from the 1999 census. Successive principal components analyses are used to select variables and create the index. Both metropolitan area-specific and global indices are tested and compared. Socioeconomic categories are drawn with hierarchical clustering as a reference to determine "optimal" thresholds able to create categories along a one-dimensional index.

Results: Among the twenty variables finally selected in the index, 15 are common to the three metropolitan areas. The index explains at least 57% of the variance of these variables in each metropolitan area, with a contribution of more than 80% of the 15 common variables.

Conclusions: The proposed procedure is statistically justified and robust. It can be applied to multiple geographical areas or socioeconomic variables and provides meaningful information to public health bodies. We highlight the importance of the classification method. We propose an R package in order to use this procedure.

Keywords: Socioeconomic status, Multidimensional index, Principal component analysis, Hierarchical classification, Small-area analysis

Social health inequalities are well documented in the epidemiological literature. Studies show that, for a wide array of health outcomes (infant mortality and pregnancy [1-3], cardiovascular and respiratory diseases [4-8], mental health [9,10], etc.), the burden of disease is different between deprived and affluent populations [11]. Most studies on social health inequalities use personal socioeconomic characteristics [1,6,7] which are often difficult and long to obtain, especially if an individual-based epidemiological study has to be set in order to collect them. Although contextual data cannot be used and interpreted as individual data (due among other issues to the ecological fallacy), it is easier to retrieve aggregate data from existing databases. Further, when the spatial units are small, as in our case, the ecological bias is reduced [12]. Besides, it is in some cases relevant, very convenient or even necessary to use this aggregate socioeconomic data as an alternative source of information for public health research [2,5,8-10,13-15].

Moreover, even when personal information is available, studies have shown that measures of neighborhood socioeconomic status (SES) explain significant variations in health status, even after adjustment for individual socioeconomic characteristics, suggesting that neighborhood SES may be by itself a risk factor [4,16,17].

* Correspondence: benoit.lalloue@ehesp.fr
[1]EHESP Rennes, Sorbonne Paris Cité, Rennes, France
[2]Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085, Rennes, France
Full list of author information is available at the end of the article

Neighborhood SES is a complex concept involving many aspects, such as employment, income, education, housing and social bonds [18-22]. Several studies have used only one variable to represent SES, a limitation that makes it difficult to take account of the multiple dimensions of neighborhood SES. Another possibility is to study the one-by-one association (for instance with simple regression models) between the outcome of interest and different variables simultaneously with the purpose to identify if a particular component of the SES is particularly associated with the outcome. However, this comparison between associations is not trivial (for instance comparison between non nested models), especially when variables are correlated, and should be carefully conducted in order to obtain sound conclusions. Eventually, it is possible to include several socioeconomic variables in the same model but this may lead to statistical problems when dealing with multicollinearity and the large number of parameters to be estimated.

To overcome these problems, a socioeconomic composite index may be set up at the level of a neighborhood, that may summarize the many aspects that encompasse the concept of SES with a large variety of possible techniques: additive scores with different weighting approaches (Z-score, experts weightings) [22-27], principal component analysis or factor analysis [22,26,28-31], spatial or GIS-based analyses [32,33], or others methods [34,35]. This kind of index can then be used for different purposes such as reveal the existence of social health inequalities through an epidemiological study , giving an overview of the situation for decision makers, or identify particular extreme areas where it is needed to focus public action (social planning, urban planning,. . .). Townsend [36] and Carstairs [24] indices, which are commonly used in various countries, cover different topics. Because they are based on only 4 variables, these indices may not always give a comprehensive representation of SES that can be used by public health bodies, at local or national level, in order to determine where action might be justified and effective. Indices that incorporate a greater number of variables from different dimensions may be more appropriate for this purpose. Moreover, these indices or Jarman's UPA[25], were built 20 years ago and may not take into account possible modifications that occurred along time and across countries in the definitions of the variables or in the evolution of how these variables may contribute to the SES. Utilization of these indices across different countries may also be difficult due to cultural or historical distinctions (e.g. "social classes", when defined, can have very different definitions; also, "proportion of households not owner of their house" vary considerably across countries); comparisons within countries may also be hampered by demographic and urban policy factors (e.g. the "proportion of households not owner of a car" highly depend on the availability of public transport and sharply contrast central urban areas and peri-urban or rural areas). Moreover, regardless of the creation procedure, the interpretation of the variables included in an index can be very different according to the implementation area. For instance, variables like ratio of individual houses or proportion of farmers do not have the same interpretation from a SES point of view in urban or rural areas. The interpretation of each variable included in an index must of course be done according to the context. In this setting, it could be helpful to have a versatile procedure which would allow taking into account these changes and selecting variables among a given set rather than fixing a mandatory list of variables.

A rigorous methodological approach is required to ensure that the index is statistically well founded and provides a good approximation of SES. Principal component analysis appears to be particularly suitable for developing composite indices because this statistical method creates non-correlated linear combinations of the variables with maximal variance, which allows the best contrast between statistical units. Furthermore, in ecologic epidemiological studies, mapping is a crucial step for showing the spatial distribution of deprivation in public health studies. While discretization of quantitative variables is frequently used for this purpose, mainly by using quantiles, it is an arbitrary technique which relies on the categories sample sizes rather than on similarities between units. Yet several other methods, such as hierarchical classification, are available to create homogeneous categories of similar spatial units.

In this context, this study presents a procedure based on statistical criteria and justification for selecting socioeconomic variables in order to create a neighborhood socioeconomic index meant to provide meaningful information to public health bodies and allow epidemiological assessment of social health inequalities. In view to assess its ability to be generalized, the procedure was applied in three contrasted French metropolitan areas to create both area-specific and global socioeconomic indices. Eventually, an R package was created in order to give an easy way to implement the procedure in a variety of contexts. This package contains the basic functions needed to run the procedure, obtain the corresponding SES index and create categories.

Material and methods

Study setting and small area level

The study was carried out in three large metropolitan areas in France: Lille (Nord Pas de Calais region, northern France), Lyon (Rhône-Alpes region, central and eastern France) and Marseille (Provence-Alpes-Côte d'Azur region, south eastern France) with a total population of around 3.8 million. These urban areas are the three largest in France after Paris. They differ in some important sociodemographic features. For instance, the Lille Métropole has a higher population of people under age 25, more blue-collar workers and individual houses than the other

cities; Grand Lyon has a higher rate of foreigners and white collar workers, and fewer people with no qualifications; while the Aix-Marseille urban unit has a higher rate of single parent families, higher unemployment, a lower rate of people with steady jobs and less social housing (see in Additional file 1 the detailed socioeconomic characteristics of the metropolitan areas).

The statistical units were the sub-municipal French census block groups (called IRIS for "Îlots Regroupés pour l'Information Statistique") defined by the National Institute of Statistics and Economic Studies (INSEE). These are the smallest units for which socioeconomic and demographic information is available from the French national census (Table 1). These units have an average of 2,000 inhabitants and are constructed in collaboration with local actors (municipalities and communities) to be as homogeneous as possible in terms of socio-demographic characteristics and land use. They also take account of physical obstacles that may break up urban landscapes, such as arterial roads, green spaces, bodies of water, and must have unambiguously identifiable contours, stable over time. Census block groups (BGs) are divided into three distinct categories in order to take into account some special cases: 1) "housing" BGs represent the vast majority of BGs (92% for all France in 2008) and count generally between 1800 and 5000 inhabitants; 2) "activity" BGs include at least 1,000 employees and at least twice as many employees as residents (e.g. industrial estates or business districts); and 3) "miscellaneous" BGs are specific wide areas sparsely populated (leisure parks, port areas, forest, etc.). As activity and miscellaneous BGs have some particular profiles due to the way they are defined, this study only considered housing BGs for the creation of the socioeconomic index. Housing BGs were treated in our statistical analysis as active units while activity and miscellaneous BGs were treated as supplementary units (meaning that they were not part of the construction of the socioeconomic index but will have an index value).

Socioeconomic data

Socioeconomic data were taken from the 1999 national census (provided by the INSEE) and provided counts of population, households and residences at BG scale covering all the social, economic and demographic aspects. Median income (for 2001) at the BG scale was taken from another national study also provided by the INSEE ("Revenus fiscaux des ménages", INSEE – DGI). Using this raw data, 48 variables were defined at the BG scale based on the INSEE definitions. These variables were chosen to be representative of the theoretical concepts of SES and in line with the variables most often used in the literature [21,22,24-31,33,36,37]. We also introduced some variables which were not totally part of the SES concept but that could be considered as linked with it (and could also give insight about the neighborhood environment). For instance we included the proportion of people over the age of 65, which could reflect the proportion of retired people; or the proportion of people who have moved from their municipality since the last census which could give indications about the "residential instability" of the neighborhood population.

All variables were associated with family structure, household type, immigration status, mobility, employment, income, education and housing (a detailed list of these variables can be found in Table 2). The aim of introducing a spectrum of variables broader than just the variables known to be indicators of the SES was to offer the possibility to examine the utility of taking into account some "proxies" and also to have a data driven approach in order to determine the variables which maximize the index's variance.

Some of the variables were intentionally redundant and represented the same notion, in view to determine which best represented this notion. There were two such groups: 7 variables of unemployment (ID 12 to ID 18 in Table 2) and 3 variables of labor force (ID 9 to ID 11 in Table 2). There were an unexpectedly high number of missing values for median income (see Table 1) but, willing to keep this variable in the analysis, we filled missing values with the average value of the adjacent BGs.

Creation of the socioeconomic index

The socioeconomic index was created by improving and expanding a procedure previously developed by our

Table 1 Description of the three study urban areas

	Lille Métropole	Grand Lyon	Aix-Marseille
Population in 1999 (inhabitants)	1,091,438	1,193,384	1,349,772
Population in 2007 (inhabitants)	1,106,885	1,260,348	1,434,845
Number of municipalities	85	58	38
Number of census block groups	506	510	630
Number of housing blocks (% of census block groups)	475 (94%)	465 (91%)	563 (89%)
Number of census block groups with missing median income (% of census blocks)	119 (24%)	95 (19%)	106 (17%)
Area (km^2)	611.45	527.15	1289.59

Table 2 Description of the first selection of 48 socioeconomic variables at the census block group scale

Unless stated otherwise, variables are proportions expressed in %		Var. Id.
Family and household	People under the age of 25 in the total population	1
	People over the age of 65 in the total population	2
	People living outside the household (boarder students, soldier in garrison, people in jail, people in nursing home or in hospital, etc.) in the total population	3
	SINGLE-PARENT FAMILIES IN THE TOTAL POPULATION	4
	Householders living alone in the total population	5
Immigration and mobility	FOREIGN PEOPLE IN THE TOTAL POPULATION	6
	FOREIGN IMMIGRANTS (SINCE THE LAST CENSUS) IN THE TOTAL POPULATION	7
	People who have moved from their municipality since the last census in the total population	8
Employment and income	*People in the labor force in the total population* [a]	9
	Men in the labor force in the total male population [a]	10
	Women in the labor force in the total female population [a]	11
	Unemployed people in the labor force [b]	12
	Unemployed foreigners in the labor force [b]	13
	Unemployed people in the 15-24 years old labor force [b]	14
	Over 50 years old unemployed people in the labor force [b]	15
	Unemployed people in the male labor force [b]	16
	Unemployed people in the female labor force [b]	17
	People unemployed for more than 1 year in the labor force [b]	18
	SELF-EMPLOYED (INDEPENDENT WORKERS, EMPLOYERS, ETC.) IN THE LABOR FORCE	19
	PEOPLE WITH UNSTABLE JOBS IN THE LABOR FORCE (APPRENTICES, TRAINEES, TEMPORARY JOBS, ETC.)	20
	PEOPLE WITH STEADY JOBS IN THE LABOR FORCE	21
	Farmers in the labor force	22
	Managers in the labor force	23
	Blue-collar workers in the labor force	24
	MEDIAN INCOME PER CONSUMPTION UNIT (IN EUROS PER YEAR)[c]	25
Education	People 6-15 years old attending school in the 6-15 years old population	26
	PEOPLE WITH NO SCHOOL GRADUATION (AND NOT STUDYING) IN THE 15 YEARS OLD AND MORE POPULATION	27
	PEOPLE WITH BASIC OR INTERMEDIATE GENERAL OR VOCATION QUALIFICATIONS (AND NOT STUDYING) IN THE 15 YEARS OLD AND MORE POPULATION	28
	PEOPLE WITH GENERAL OR VOCATIONAL MATURITY CERTIFICATES (AND NOT STUDYING) IN THE 15 YEARS OLD AND MORE POPULATION	29
	People with at least a lower tertiary education (and not studying) in the 15 years old and more population	30
	People with a higher educational degree (and not studying) in the 15 years old and more population	31
	Students in the 15 years old and more population	32
Housing	Individual houses in the main residences	33
	Multiple dwelling units in the main residences	34
	NON-OWNER-OCCUPIED IN THE MAIN RESIDENCES	35
	Subsidized housing in the main residences	36
	Main residences built before 1968	37
	Main residences built after 1990	38
	Main residences less than 40 m^2	39
	Main residences larger than 150 m^2	40

Table 2 Description of the first selection of 48 socioeconomic variables at the census block group scale *(Continued)*

	Main residences without bathtub or shower	41
	Main residences without toilet	42
	Main residences without heating	43
	Main residences with a parking space (garage or other)	44
	MAIN RESIDENCES WITH MORE THAN ONE PERSON PER ROOM	45
	AVERAGE NUMBER OF PEOPLE PER ROOM[c]	46
	HOUSEHOLDS WITHOUT A CAR	47
	HOUSEHOLDS WITH 2 OR MORE CARS	48

[a]Redundant group "labor force".
[b]Redundant group "unemployment".
[c]Not a proportion.
UPPERCASE : variables selected commonly for each metropolitan areas and global analysis.
Italic: variables selected during the "reduction of redundant groups" step for the global analysis.

team [38]. The three steps (Figure 1) described below were used:

Step 1 - Study of the redundant variables (in our case variables ID 9-11 and ID 12-18 aforementioned). To avoid issues due to redundant covariates (correlation over 0.8 for most of them in all the study areas) one variable was selected from each group by applying principal component analysis (PCA, see Additional file 2) to each of the two groups of redundant variables. The first component was a good representation of the group of variables, strongly correlated with all of the variables (if not and if a variable was not well represented by the first component, - a situation that did not happen in our case - this variable should actually not have been part of the redundant group and should not have been included in it). However, to ease interpretation, the variable with the largest correlation with the first component was selected. Reducing the two redundant groups to one variable each reduced, in our case, the number of variables to 40 (the two selected variables at this step are shown in italic font in Table 2).

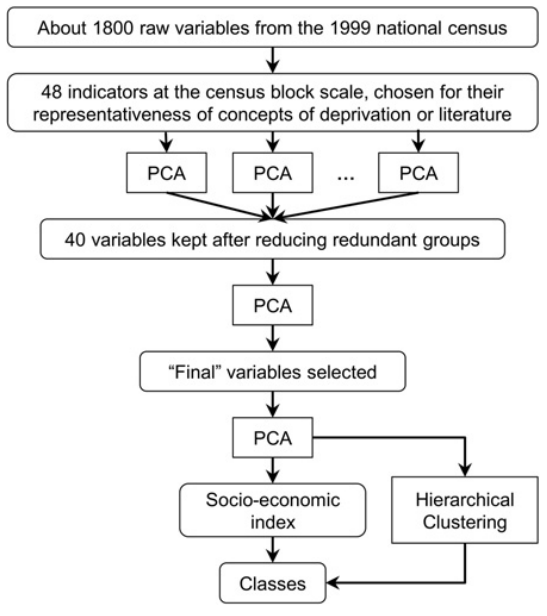


Figure 1 Steps of the socioeconomic index creation.

Step 2 - Selection of the variables. PCA on these 40 variables (i.e. after selection of one variable per group of redundant variables) was used to select the variables with a contribution to the first component larger than the average one, i.e. variables that were best correlated with the first component.

Step 3-Construction of the final index. A final PCA was carried out including the variables selected in step 2. Provided that the first component of this PCA could be interpreted (according to the meaning of the variables in the given context) as a "SES component" (which was expected giving the variables selected and confirmed *a posteriori*), it was used to calculate the socioeconomic index as the reduced first component. This normalization gave an index with mean 0 and standard deviation 1. Since the purpose of the whole procedure was to create a single index, the second and subsequent components of the final PCA were not used (in practice, the variance explained by the second component was far below that explained by the first component and there was no clear interpretation of it).

The choice of PCA as the main technique in the procedure was done in order to use as little constraints and hypotheses as possible, as well as to keep a strongly data driven approach. This approach is not based on a model set a priori, like Factor Analysis. Moreover, it is known that a PCA where the last eigenvalues are near and close to 0 gives results very similar to those of a Factor Analysis.

This procedure was applied to each metropolitan area independently (giving socioeconomic indices specific to each area) and to the three urban areas altogether. This produced four different indices that, as a result, could be compared. Step 1 was kept in the procedure and applied to each metropolitan area because differences between them in their socioeconomic make-up might lead to different choices.

Hierarchical clustering and optimal thresholds

Socioeconomic categories were created so that the index could be used as an explanatory variable to determine possible non-linear relationships in a variety of applications, and be used for mapping.

Defining socioeconomic index quantiles is a common technique but it is sometimes unsatisfactory. Indeed, classes created with quantiles are only based on the number of units and therefore may not correctly classify units according to their similarity if they are not distributed homogeneously. This can lead to merging in the same class very different groups of units, or to split into two classes a homogeneous group.

Hierarchical clustering (HC) is frequently used after PCA [39] in data mining to create meaningful categories. Given a set of p variables measured on n elements, each element is represented as a point in $\mathbb{R}^p$. A distance between elements d (usually the Euclidian distance) and a distance between categories Δ (based on d) are defined. HC algorithm creates a hierarchy of categories step by step by merging at each step the two categories which are the nearest according to Δ. When Δ is a particular distance (the Ward's distance), this algorithm allows to obtain categories homogeneous in their composition and heterogeneous between them (i.e. with a maximum between-categories inertia). The most appropriate partition is then selected from the hierarchy of categories. More methodological details about HC are available in Appendix 2.

However, HC is a multidimensional technique which uses several components of a PCA (often 5 or 10). Now, we wanted here to create the categories from a one-dimensional index (it is also possible, but not for the same purpose, to keep directly the categories created by HC in order to have a qualitative index). Then, we used HC as a reference to determine "optimal" thresholds able to construct a one-dimensional classification as close as possible to the HC. There were two possible cases, depending on the number of categories:

- either the categories constructed with HC were not distributed along the first principal axis of the PCA (the second and subsequent axes affecting classification) in which case it was not possible to determine thresholds along the first axis that would be able to correctly approximate HC categories. The index could not be used by itself and the number of categories was therefore reduced.
- Or the categories were distributed along the first component of the PCA (i.e. our socioeconomic index). In this case, optimal thresholds were determined using a simple iterative algorithm: at each step, categories were defined with new thresholds along the index values and the concordance rate between this classification and the clustering using HC were calculated. Only values with the best concordance percentage were kept. It created socioeconomic categories using the socioeconomic index by itself.

Comparison of indices and classifications

Pearson's coefficient of correlation was used to compare the area-specific indices to the global one, which encompasses the 3 metropolitan areas, and also the Carstairs' and Tonwsend indices with ours.

Carstairs' index [24] was constructed as the sum of the standardized proportions of total unemployment, of households without a car, of households with more than one person per room, and of blue-collar workers (since French census do not use "social classes"). Townsend's index [36] was constructed as the sum of the standardized log-proportions of total unemployment and households with more than one person per room, and the proportions of households without a car, and of non-owner-occupied main residences.

The concordance percentage was used to compare pairs of classifications. This is the percentage of BGs in the same class in both classifications (the diagonal of the confusion matrix). The R software [40], with the FactoMineR package [39] and the SesIndexCreatoR package, was used to create the indices and clustering, to determine thresholds and draw comparisons.

R package

Since the procedure described here is more complex than for some other SES indices, we specifically developed the SesIndexCreatoR package. The version 1.0 of this package (currently freely available on the website of the Equit'Area project: http://www.equitarea.org/documents/packages_1.0-0/) contains the basic functions needed to run the procedure (in its entirety or only in some steps) and to obtain the corresponding SES index. The user may also create categories of this index with different methods (hierarchical clustering with or without k-nearest neighbors, quantiles, or intervals). We project to extend the package in the future and among other improvements we foresee to add tools to help the interpretation of these categories and the visualization of the results.

Results

Constructing indices, selecting variables and determining contributions

The four socioeconomic indices were built as the first component of PCA using the data for each urban area separately and for an overall analysis of the combined sets of data. In each case, this first component of the PCA was positively correlated with variables of low SES (unemployment, single-parent families, overcrowding, etc.) and negatively correlated with variables of high SES (income, steady jobs, high level of education, etc.); therefore, it was interpreted as a true SES component (see Figure 2). It always explained a large proportion of the total variance (Table 3). By contrast, the second component had never a clear interpretation and explained less than 17% of the total variance.

Fifteen of the 20 variables selected by the procedure (step 1 to 3) were common to all four indices and accounted for more than 77% of the construction of the first component (see below). This result showed that the procedure was robust and that, despite the substantial socio-demographic differences between the urban areas, the same common variables could explain a large part of the socioeconomic variability. Now, the ranks of the contributions of the variables to the index were different in each urban area (Additional file 3) although none was far from the overall average contribution.

For Lille Métropole, the procedure selected 21 variables for the index (for Grand Lyon, Marseille urban area and global analysis it selected 20, 20 and 19, respectively). The first component of the final PCA explained 61% (resp. 58%, 57% and 57%) of the total variance while the second 12% (resp. 17%, 15% and 11%). The variables common to all four indices contributed 80% (resp. 84%, 78% and 88%) to the index. In each metropolitan area, local experts associated with the project confirmed that our SES index globally well-represent the socioeconomic true profile of the neighborhoods.

Comparisons between indices

The indices in each metropolitan area were compared with the global index restricted respectively to the BGs of each area (the global index was constructed on the BGs of the three areas altogether but, to allow comparison with the area-specific index, only BGs of this area were considered), as well as with those proposed by Carstairs and Townsend.

In general, all the correlations between the area-specific and the overall index restricted to the BGs of each metropolitan area were above 0.9 (Table 4) with a clear linear association (Additional files 4, 5 and 6). There were very good correlations between our indices (constructed either for each urban area or for the three areas altogether) and the Carstairs and Townsend indices (always larger than 0.91) suggesting that the socioeconomic dimension measured by our index is very close to that measured by the well-known and often used Carstairs and Townsend indices.

Comparisons between classifications

The initial number of categories we tested for classification was five because it is a usual number of categories used in spatial epidemiology, especially for mapping. For all four indices, the categories obtained through HC depended both on the first and second axes of the PCA and the categories were not distributed solely

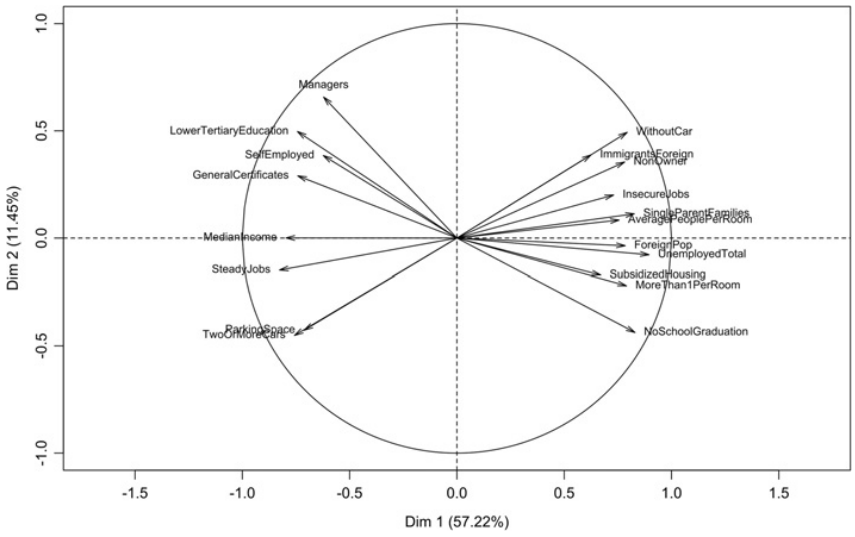


Figure 2 Circle of correlation for final step, global analysis.

Table 3 Percentage of variance explained by the two first components of the final PCA, by area

	Lille Métropole	Grand Lyon	Marseille urban unit	Global
1st component	60.73%	57.79%	57.29%	57.22%
2nd component	12.13%	16.71%	14.66%	11.45%

along the first component. This observation suggests that 5 categories were too many should the objective be to use only the first component (i.e. our index). The analysis showed that classification with quintiles was not optimal, since the hierarchical clustering gave very different results. The number of categories was, therefore, reduced until categories could be constructed only using the first axis of the final PCA. In each case, the largest number of categories was three. This number creates the optimal classification according to the HC. Average values of the common variables for the global analysis can be seen in Additional file 7.

Concordance rates between the different techniques are shown in Table 5. Very low concordance was found for the HC with 5 categories and quintile classifications, with less than 50% of similar classifications for the three metropolitan areas and about 60% for the metropolitan areas taken together. HC was unable to construct 5 categories using only the first PCA axis. The concordance rate between HC with 3 categories and terciles was between 69% and 78%. This could be explained by the smaller number of categories. Even such concordance rates mean that about one quarter of the BGs had a different class across the two classification methods. When comparing the SES index categories (created by HC or by optimal threshold) with Carstairs' and Townsend's indices categorized through quantiles, the concordance rates remained low.

By contrast, the concordance rates between HC in 3 categories and optimal thresholds were between 93% and 97%, confirming that the classification in three categories was fully along the first axis. Additional files 8, 9 and 10 show for each metropolitan area maps of the three socioeconomic categories created either by tiertiles or optimal thresholds, and the range of the SES index for each category (the higher the index, the most deprived the area)..

Table 4 Correlation coefficients between indices

	Lille Métropole	Grand Lyon	Marseille urban unit	Global
Global [a]	0.99	1	0.99	/
Carstairs	0.92	0.96	0.91	0.94
Townsend	0.98	0.94	0.96	0.96

[a]When comparing the global index with a city index, global index is restricted to that particular city's census block groups.

Discussion

This study developed a rigorous statistical procedure to create socioeconomic indices in urban contexts, improving and extending a previous work.[38] The procedure was applied and validated on three different French urban areas and proved its robustness in different socio-demographic settings. An R package was developed in order to help applying this procedure in other contexts.

As for most studies developing new methodologies to construct a neighborhood socioeconomic index, the preliminary selection of variables was based on a literature review [10,15,26,27,29]. The social and material deprivation index developed by Pampalon et al [28] included people without qualifications, employment ratio, average income, individuals living alone, individuals divorced, separated or widowed, and single parent families. They chose these variables according to four criteria: well documented health links, variables previously used as "geographic proxies" in social health inequality studies, variables belonging to the material or social dimension of deprivation and the availability of data for their study area. Carstairs and Townsend followed a similar procedure for selecting 4 variables characterizing neighborhood deprivation in their indices. However, this approach was only a preliminary step in the construction of our index. One originality of our procedure lays in selecting the final variables for the index by usage of data mining techniques rather than only information gleaned from a literature review, allowing to discard part of the subjectivity that may influence the choice of the variables. This data driven approach allows the data "speak by themself". Although it was what we expected, it was not sure, before the PCA was implemented that the first component would be a good socioeconomic index. This appeared *a posteriori* as the PCA explored the data and revealed their underlying structure.

About 20 variables, a number not defined *a priori*, were selected for each metropolitan area, encompassing the various domains of SES. This allowed to determine the common determinants of SES in the various areas and also to select determinants which are more specific in each area. The larger number of variables compared with other indices gives room for a finer spatial description of SES and of specific characteristics of each metropolitan area, providing information which public health bodies might find helpful in determining key targets for local actions. Indeed, once the index is constructed and used to identify BGs with the lowest SES, it is possible to return to the variables that compose the index in order to see which ones could be a leverage for action, a property that more simple indexes lack. Using this method (use such an index, in a quantitative or qualitative way, to identify lowest SES areas and then go back to the individual variables to have more details) in an epidemiological study to describe the spatial distribution of some disease or cause of mortality in

Table 5 Concordance rates between different clustering techniques and between indexes

	HC (5) [a] vs. quintiles [b]	HC (3) [a] vs. tiertiles [b]	HC (3) [a] vs. optimal thresholds [b]	Optimal thresholds vs. tiertiles [b]	Carstairs		Townsend	
					HC (5) [a] vs. quintiles [b]	Optimal thresholds vs. tiertiles [b]	HC (5) [a] vs. quintiles [b]	Optimal thresholds vs. tiertiles [b]
Lille Métropole	41%	78%	98%	79%	38%	70%	42%	78%
Grand Lyon	48%	74%	93%	78%	47%	77%	40%	75%
Marseille urban unit	48%	69%	97%	67%	51%	67%	50%	69%
Global	63%	71%	97%	72%	57%	70%	55%	71%

[a]Hierarchical Clustering using Principal Components (in parenthesis, the number of categories chosen).
[b]Concordance rate (percent of census block groups categorized into the same class using the two different clustering schemes).

a metropolitan area will not only allow to flag communities where the risk is highest, but will also provide information on the social and economic characteristics of these communities upon which appropriate and focused preventive policies can be devised and implemented.

The large number of common variables (15 of the 20 variables) across the metropolitan areas shows the stability of the results and the good representation of the underlying concept of SES conveyed by the index. These variables reveal the common determinants of SES in different French metropolitan areas, at BG level, which is the smallest administrative unit for which census data is available. The specific SES patterns in each area can be assessed in two different ways: through the variables which are specific to each area, and through the relative contribution of each variable to the final index. As a result, the procedure proposed in this study can be used alternatively to build a city-specific index which can be applied locally, for instance to determine priority BGs for local action, or a global index to compare a set of cities with the same metric.

However, one should remember that data and indices used here are area-based and not person-based. Indeed, although BGs are constructed in order to be as homogeneous as possible, there is still individual variability within them which cannot be assessed by aggregated data. Therefore, as it is now well-known, inference at the individual scale from indices created at the BG scale can be tricky due to the ecological fallacy. SES indices presented here are neighborhood SES indices and should be used as a way to assess the contextual socioeconomic setting in which people live rather than a way to approximate the individual SES.

When socioeconomic indices were first constructed, categories were delineated to show the spatial distribution of SES on maps and to investigate the existence of non-linear social relationships with some outcome of interest. So far, to our knowledge, most of the studies classifying deprivation scales have used quantiles [2,10,13,15,27,29,33] without questioning the validity of this classification method from a statistical point of view. This simple approach should be used with caution; our study suggests that it might put dissimilar geographical units in the same class and separate similar units, according to HC.

Using HC, the first dimension alone of the final PCA was not sufficient to create 5 socioeconomic categories. Although we could have kept the results of the HC as a qualitative index, this would have contradicted with our aim to have a one-dimensional index. In this study, but without possible generalization to other data, it was preferable to use a 3-categories classification built only with the first component of the final PCA.

Despite its statistical justification, this study has some limitations. Some are induced by the very nature of an index. Since indices are composite syntheses of several variables, they have no unit. This can reduce the interpretability of their application, especially regression models, the meaning of an increase or decrease of one unit of the SES index being difficult to express. From a public policy point of view, an index alone cannot give indications on how to operate to change the situation. Although the indices created by the procedure we propose share these limitations, we think they are interesting as first indicators of 'global' neighborhood SES and as a synthetic tool to point out the situation to policy makers. Eventually, one may return, as aforementioned, to the variables composing the index to have a better insight of the actual situation of the identified neighborhoods and the variables that most contribute to this signal.

Secondly, median income had to be estimated where the data was missing. Because BGs with incomplete information on median income were a minority (maximum 24% for the Lille metropolitan area) and because only one variable among the 20 used in the indices had such missing data, incompleteness has probably little effect. A perspective for improvement could be to use more advanced techniques to handle missing data.

Thirdly, utilization of a large amount of data requires preparation and calculation before applying the procedure, which is time consuming. It also calls for technical know-how. This procedure is clearly more complex than number of other indices. We think this is the price to pay for a deeper analysis of SES and its determinants and a more

detailed interpretation of the results. While our index showed a high correlation with the Carstairs and Townsend indices, we think it allows more in depth analysis, when needed, and overcomes some of the limitations faced by between and within countries comparisons due to the low number and the nature of the variables than compose these well-known indices. Similar studies in other countries that allow usage of detailed socioeconomic information at BG levels would help assess the robustness of the procedure in other social contexts.

Fourthly, HC has no criteria regarding the size of the categories and so it can yield categories with very different sizes, which can be a limitation when linking their distribution with other attributes such as the prevalence of some health condition or of some exposure factor.

As a summary, a major strength of the procedure presented in this article is its versatility: it is not restricted to a particular set of data or type of study, and can be used for a large variety of contexts such as social epidemiology, environmental justice assessment, public health studies or urban and social planning. The application of this procedure on three large metropolitan areas shows high correlations with well-known indices like Townsend's and Carstairs', which appears to confirm that the created index represents the same socioeconomic notion. Although this procedure is more complicated than these other methods to create a SES index, the variables included in the final SES index allows a wider representation of the dimensions of SES, both to identify the best variables to distinguish BGs at the metropolitan area scale and to have better information on the particularities of the BGs. Then, it allows finer analysis of key determinants of health inequalities and reflection on local policies that would aim to cope with these inequalities. Another innovation in this study is the use of HC to constitute SES categories and compare them to the classically used quantiles. This approach allows having categories with more homogeneous compositions and which can consequently increase contrasts between them. Finally, we provide an R package able to reproduce the procedure easily.In conclusion, this procedure can be used to produce a SES index with a strong statistical basis and great scope for interpretation and relevance to public health bodies. The set of selected variables had a high proportion of common determinants of SES; they could also identify some features more specific to each area. Comparison of clustering methods showed that care should be taken to derive homogeneous categories.

Additional files

Additional file 1: Base Socioeconomic Characteristics of the Three Study Urban Areas.

Additional file 2: Principal component analysis and Hierarchical clustering.

Additional file 3: Correlations and Contributions of Variables to the First Component and Variance Explained by the First Component, According to the Study Area.

Additional file 4: Plot of city SES index vs. global index restricted to each city (housing census block groups only).

Additional file 5: Plot of SES index vs. Carstairs' index, according to the study area (housing census block groups only).

Additional file 6: Plot of SES index vs. Townsend's index, according to the study area (housing census block groups only).

Additional file 7: Average Values of the Common Variables per Category Created With HC for the Global Analysis.

Additional file 8: Maps of the socioeconomic index for Lille Metropole, in three categories by tiertiles or optimal thresholds.

Additional file 9: Maps of the socioeconomic index for Grand Lyon, in three categories by tiertiles or optimal thresholds.

Additional file 10: Maps of the socioeconomic index for Aix-Marseille urban area, in three categories by tiertiles or optimal thresholds.

Abbreviations

BG: Census block group; HC: Hierarchical clustering; PCA: Principal component analysis; SES: Socio-economic status.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

BL has contributed to the creation of the method, implemented the procedure, created the R package, performed and interpreted the application, drafted the article and conducted the literature review. JMM has contributed to the creation of the method, provided statistical rigor, contributed to the interpretation of the results and helped to draft the manuscript. CP has collected socioeconomic data and helped to draft the manuscript. WK has constructed some socioeconomic variables, filled the missing data for median income and helped to draft the manuscrip. NLM has given insight and expertise for the implementation and creation of the R package and helped to draft the manuscript. DZN guarantees quality assurance and helped to draft the manuscript. SD, head of Project Equit'Area research examining the role of environmental exposures on health inequalities, has followed up the general labor, has contributed to the definition of the method, interpretation of results, writing section and its finalization. All authors read and approved the final manuscript.

Acknowledgments

This work and the Equit'Area project are supported by the French National Research Agency (ANR, contract-2010-PRSP-002-01) and the EHESP School of Public Health. This research was also jointly supported by the Direction Générale de la Santé (DGS), the Caisse Nationale d'Assurance Maladie des Travailleurs Salariés (CNAMTS), the Régime Social des Indépendants (RSI), the Caisse Nationale de Solidarité pour l'Autonomie (CNSA), the Mission Recherche de la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (MiRe-DREES) and l'Institut national de prévention et de promotion de la santé (Inpes), under the research call launched by the French Institute of Public Health Research (IReSP) in 2010.

Author details

[1]EHESP Rennes, Sorbonne Paris Cité, Rennes, France. [2]Inserm, UMR IRSET Institut de recherche sur la santé l'environnement et le travail - 1085, Rennes, France. [3]Lorraine University, CNRS, INRIA UMR 7502, Institut Elie Cartan, Lorraine, France. [4]UMR936 INSERM, Université de Rennes 1, Rennes, France. [5]Lorraine University, Medical School, Lorraine, France.

Received: 20 December 2012 Accepted: 17 March 2013
Published: 28 March 2013

References

1. Arntzen A, Samuelsen SO, Bakketeig LS, Stoltenberg C: Socioeconomic status and risk of infant death. A population-based study of trends in Norway, 1967-1998. *Int J Epidemiol* 2004, **33**:279–288.

2. Krieger N, Chen JT, Waterman PD, Soobader M-J, Subramanian SV, Carson R: **Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US).** *J Epidemiol Community Health* 2003, **57:**186–199.
3. Singh GK, Kogan MD: **Persistent socioeconomic disparities in infant, neonatal, and postneonatal mortality rates in the United States, 1969-2001.** *Pediatrics* 2007, **119:**e928–e939.
4. Chaix B, Rosvall M, Merlo J: **Recent increase of neighborhood socioeconomic effects on ischemic heart disease mortality: a multilevel survival analysis of two large Swedish cohorts.** *Am J Epidemiol* 2007, **165:**22–26.
5. Deguen S, Lalloué B, Bard D, Havard S, Arveiler D, Zmirou-Navier D: **A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex: a Bayesian modeling approach.** *Epidemiology* 2010, **21:**459–466.
6. Ellison-Loschmann L, Sunyer J, Plana E, Pearce N, Zock J-P, Jarvis D, Janson C, Antó JM, Kogevinas M: **Socioeconomic status, asthma and chronic bronchitis in a large community-based study.** *Eur Respir J* 2007, **29:**897–905.
7. Prescott E, Godtfredsen N, Vestbo J, Osler M: **Social position and mortality from respiratory diseases in males and females.** *Eur Respir J* 2003, **21:**821–826.
8. Tassone EC, Waller LA, Casper ML: **Small-Area Racial Disparity in Stroke Mortality.** *Epidemiology* 2009, **20:**234–241.
9. Curtis S, Copeland A, Fagg J, Congdon P, Almog M, Fitzpatrick J: **The ecological relationship between deprivation, social isolation and rates of hospital admission for acute psychiatric care: a comparison of London and New York City.** *Health Place* 2006, **12:**19–37.
10. Tello JE, Jones J, Bonizzato P, Mazzi M, Amaddeo F, Tansella M: **A census-based socio-economic status (SES) index as a tool to examine the relationship between mental health services use and deprivation.** *Soc Sci Med* 2005, **61:**2096–2105.
11. Marmot M: **Social determinants of health inequalities.** *Lancet* 2005, **365:**1099–1104.
12. Elliott P, Wakefield J, Best N, Briggs D: *Spatial epidemiology: methods and applications.* Oxford; New York: Oxford University Press; 2001.
13. Barceló MA, Saez M, Saurina C: **Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain.** *Sci Total Environ* 2009, **407:**5501–5523.
14. Borrell C, Marí-Dell'olmo M, Serral G, Martínez-Beneito M, Gotsens M: **Inequalities in mortality in small areas of eleven Spanish cities (the multicenter MEDEA project).** *Health Place* 2010, **16:**703–711.
15. Singh GK: **Area deprivation and widening inequalities in US mortality, 1969-1998.** *Am J Public Health* 2003, **93:**1137–1143.
16. Zeka A, Melly SJ, Schwartz J: **The effects of socioeconomic status and indices of physical environment on reduced birth weight and preterm births in Eastern Massachusetts.** *Environ Health* 2008, **7:**60.
17. Meijer M, Röhl J, Bloomfield K, Grittner U: **Do neighborhoods affect individual mortality? A systematic review and meta-analysis of multilevel studies.** *Soc Sci Med* 2012, **74:**1204–1212.
18. Braveman PA, Cubbin C, Egerter S, Chideya S, Marchi KS, Metzler M, Posner S: **Socioeconomic status in health research: one size does not fit all.** *JAMA* 2005, **294:**2879–2888.
19. Gordon D: **Census based deprivation indices: their weighting and validation.** *J Epidemiol Community Health* 1995, **49**(Suppl 2):S39–S44.
20. Krieger N, Williams DR, Moss NE: **Measuring social class in US public health research: concepts, methodologies, and guidelines.** *Annu Rev Public Health* 1997, **18:**341–378.
21. Galobardes B, Shaw M, Lawlor DA, Lynch JW, Davey Smith G: **Indicators of socioeconomic position (part 2).** *J Epidemiol Community Health* 2006, **60:**95–101.
22. Morris R, Carstairs V: **Which deprivation? A comparison of selected deprivation indexes.** *J Public Health Med* 1991, **13:**318–326.
23. Carstairs V: **Deprivation indices: their interpretation and use in relation to health.** *J Epidemiol Community Health* 1995, **49**(Suppl 2):S3–S8.
24. Carstairs V, Morris R: **Deprivation: explaining differences in mortality between Scotland and England and Wales.** *BMJ* 1989, **299:**886–889.
25. Jarman B: **Identification of underprivileged areas.** *Br Med J (Clin Res Ed)* 1983, **286:**1705–1709.
26. Eibner C, Sturm R: **US-based indices of area-level deprivation: results from HealthCare for Communities.** *Soc Sci Med* 2006, **62:**348–359.
27. Fukuda Y, Nakamura K, Takano T: **Higher mortality in areas of lower socioeconomic position measured by a single index of deprivation in Japan.** *Public Health* 2007, **121:**163–173.
28. Pampalon R, Hamel D, Gamache P, Raymond G: **A deprivation index for health planning in Canada.** *Chronic Dis Can* 2009, **29:**178–191.
29. Rey G, Jougla E, Fouillet A, Hémon D: **Ecological association between a deprivation index and mortality in France over the period 1997–2001: variations with spatial scale, degree of urbanicity, age, gender and cause of death.** *BMC Public Health* 2009, **9:**33.
30. Salmond C, Crampton P, Sutton F: **NZDep91: A New Zealand index of deprivation.** *Aust N Z J Public Health* 1998, **22:**835–837.
31. Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, Elo I, Burke JG, O'Campo P: **The development of a standardized neighborhood deprivation index.** *J Urban Health* 2006, **83:**1041–1062.
32. Anthopolos R, James SA, Gelfand AE, Miranda ML: **A spatial measure of neighborhood level racial isolation applied to low birthweight, preterm birth, and birthweight in North Carolina.** *Spat Spatiotemporal Epidemiol* 2011, **2:**235–246.
33. Bell N, Schuurman N, Hayes MV: **Using GIS-based methods of multicriteria analysis to construct socio-economic deprivation indices.** *Int J Health Geogr* 2007, **6:**17.
34. Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, Lang T, Launoy G: **Construction of an adaptable European transnational ecological deprivation index: the French version.** *J Epidemiol Community Health* 2012, **66**(11):982–989.
35. Marí-Dell'Olmo M, Martínez-Beneito MA, Borrell C, Zurriaga O, Nolasco A, Domínguez-Berjón MF: **Bayesian factor analysis to calculate a deprivation index and its uncertainty.** *Epidemiology* 2011, **22:**356–364.
36. Townsend P, Phillimore P, Beattie A: *Health and deprivation: inequality and the North.* London: Routledge; 1988.
37. Carstairs V, Morris R: **Deprivation and mortality: an alternative to social class?** *J Public Health* 1989, **11:**210–219.
38. Havard S, Deguen S, Bodin J, Louis K, Laurent O, Bard D: **A small-area index of socioeconomic deprivation to capture health inequalities in France.** *Soc Sci Med* 2008, **67:**2007–2016.
39. Lê S, Josse J, Husson F: **FactoMineR: An R package for multivariate analysis.** *J Stat Software* 2008, **25:**1–18.
40. R Development Core Team: *R: A language and environment for statistical computing.* Vienna Austria: R Foundation for Statistical Computing; 2011.

doi:10.1186/1475-9276-12-21
Cite this article as: Lalloué *et al.*: A statistical procedure to create a neighborhood socioeconomic index for health inequalities analysis. *International Journal for Equity in Health* 2013 **12**:21.

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

SesIndexCreatoR: An R Package for Socioeconomic Indices Computation and Visualization

Benoît Lalloué
EHESP & Lorraine University

Séverine Deguen
EHESP

Jean-Marie Monnez
Lorraine University

Cindy Padilla
EHESP

Wahida Kihal
EHESP

Denis Zmirou-Navier
EHESP & Lorraine University

Nolwenn Le Meur
EHESP

Abstract

In order to study social inequalities, indices can be used to summarize the multiple dimensions of the socioeconomic status. As a part of the Equit'Area Project, a public health program focused on social and environmental health inequalities, a statistical procedure to create (neighborhood) socioeconomic indices was developed. This procedure uses successive principal components analyses to select variables and create the index. In order to simplify the application of the procedure for non specialists, the R package **SesIndexCreatoR** was created. It allows the creation of the index with all the possible options of the procedure, the classification of the resulting index in categories using several classical methods, the visualization of the results, and the generation of automatic reports.

Keywords: socioeconomic status, multidimensional index, principal component analysis, hierarchical classification, R.

1. Introduction

When studying social inequalities, it is generally interesting to take into account the socioeconomic status (SES) of an individual, a neighborhood or a region rather than consider

only one socioeconomic variable such as educational level or income. However, socioeconomic status is a complex and multidimensional concept which encompasses many aspects such as employment, income, education, housing and social bonds. All of these aspects can themselves be represented by various variables. To synthesize and consider these different aspects, one solution is to create a SES index.

There are already many existing SES indices, especially at the neighborhood level (Jarman 1983; Morris and Carstairs 1991; Carstairs 1995; Salmond *et al.* 1998; Eibner and Sturm 2006; Messer *et al.* 2006; Bell *et al.* 2007; Fukuda *et al.* 2007; Pampalon *et al.* 2009). However, most of them use a small number of variables, combine variables with simple methods (such as Z-score) and/or select variables only from the litterature, which seems inappropriate for the purpose of the Equit'Area Project, a public health program focused on social and environmental health inequalities (`http://www.equitarea.org`), as detailed elsewhere (Lalloué *et al.* 2013). Thus, a new statistical procedure to create neighborhood socioeconomic indices was developed. Basically, this procedure does not create an index from a set of determined and precise variables, but aims to select, from a large data set, variables which will compose the SES index. It is based on several successive principal component analyses and the whole procedure is detailed in the aforementioned article. It has already been successfully used in several analyses aiming to study health or environmental inequalities (Padilla *et al.* 2013a,b).

Compared to other existing approaches to compute indices, our procedure is a slightly more complex to understand and apply, especially for non statisticians. Therefore we have implemented our model in a R (R Core Team 2013) package, named **SesIndexCreatoR**. The package is freely available on the website of the Equit'Area project and on CRAN. The purpose of this package is to give tools as simple as possible to perform the procedure while keeping the various possibilities it offered, like using different data mining methods, adding illustrative units, or performing only one step of the procedure. Moreover, once the index is created, users can display all the results of the different analyses both in text and graphical output, and generate a report summary.

In this paper we present and illustrate the use of the **SesIndexCreatoR** package for Lille agglomeration (a large French metropolitan area). For further examples we recommend reading the works by Padilla et al as mentioned above.

2. Material and methods

2.1. Data

The example data provided in the **SesIndexCreatoR** package concerns one large city in France, Lille (Nord Pas de Calais region, northern France), and some adjacent municipalities. The statistical unit is the sub-municipal French census block groups (called IRIS) defined by the National Institute of Statistics and Economic Studies (INSEE). These units have an average of 2,000 inhabitants and are constructed to be as homogeneous as possible in terms

of socio-demographic characteristics and land use. Census block groups (BGs) are divided into three distinct categories: housing, economical activity and miscellaneous. Housing BGs are the most common, economical activity BGs include at least 1,000 employees and at least twice as many employees as residents, and miscellaneous BGs are specific wide areas sparsely populated (leisure parks, port areas, forest, etc.). As activity and miscellaneous BGs have some particular profiles due to the way they are defined, they are treated in the example as illustrative units (meaning that they are not part of the procedure but will have an index value). For confidentiality and distribution reasons, the real BGs idenficators are replaced in the example data set with a simple number from 1 to 234 (which is the number of BGs of the area).

Socioeconomic data are taken from the 1999 national census (source: INSEE) and provide counts of population, households and residences at BG scale covering all the social, economic and demographic aspects. Median income at the BG scale is taken from a second database: the "Revenus fiscaux des ménages" database (source: INSEE-DGI). Using this raw data, 37 variables are defined at the BG scale based on the INSEE definitions. These variables are chosen to be representative of the theoretical concept of SES and in line with the variables most often used in the literature, or that could be considered as linked with the SES concept. All variables are related to family structure, household type, immigration status, employment, income, education and housing (more details are available in Table 1 and Table 2). Some of the variables are intentionally redundant and represent the same notion, in view to determine which best represents this notion (using the algorithm implemented in the proposed package). In our example, there are two such groups: 7 variables of unemployment and 3 variables of labor force. We also note there are an unexpectedly high number of missing values for median income but, willing to keep this variable in the analysis, we filled missing values with the average value of the adjacent BGs.

2.2. SES index creation

The SES index creation procedure is detailed in Lalloué *et al.* (2013). Basically, it follows three successive steps :

1. *Study of the redundant variables.* As already mentioned, several variables represent the same notion and we want to determine which best represented this notion. Therefore, one variable is selected for each group by applying principal component analysis (PCA) to each of the groups of redundant variables. The selected variable for each group is the one with the largest correlation with the first component of the PCA on the group.

2. *Selection of the variables.* A PCA or a multiple factor analysis (MFA) on the remaining variables (i.e., non redundant variables and variables selected in step 1) is used to select the variables with a contribution to the first component larger than the average one, i.e., variables that were best correlated with the first component. The choice of PCA or MFA depends on the willingness to give the same weight in the analysis to each domain (MFA) or not (PCA).

3. *Construction of the index.* A final PCA is carried out including the variables selected in step 2. Provided that the first component of this PCA could be interpreted as a

"SES component", it is used to calculate the socioeconomic index as the reduced first component.

Domain	*Variable name*	*Description*
BG type	`Type`	Census block group type (H: housing ; A: activity ; D: miscellaneous ; Z: one BG municipality) [c]
Family and Household	`UnderAge25`	People under the age of 25 in the total population
	`OverAge65`	People over the age of 65 in the total population
	`SingleParentFamilies`	Single-parent families in the total population
	`HouseholderAlone`	Householders living alone in the total population
Immigration	`ForeignPop`	Foreign people in the total population
Employment and income	`LabourForce`	People in the labor force in the total population [a]
	`MenLabourForce`	Men in the labor force in the total male population [a]
	`WomenLabourForce`	Women in the labor force in the total female population [a]
	`UnemployedTotal`	Unemployed people in the labor force [b]
	`UnemployedForeigners`	Unemployed foreigners in the labor force [b]
	`UnemployedAge1524`	Unemployed people in the 15-24 years old labor force [b]
	`UnemployedOverAge50`	Over 50 years old unemployed people in the labor force [b]
	`UnemployedMen`	Unemployed people in the male labor force [b]
	`UnemployedWomen`	Unemployed people in the female labor force [b]
	`UnemployedMore1Year`	People unemployed for more than 1 year in the labor force [b]
	`SelfEmployed`	Self-employed (independent workers, employers, etc.) in the labor force
	`InsecureJobs`	People with unstable jobs in the labor force (apprentices, trainees, temporary jobs, etc.)
	`SteadyJobs`	People with steady jobs in the labor force
	`MedianIncome`	Median Income per consumption unit (in euros per year) [c]

Table 1: Description of 37 socioeconomic variables available for the Lille agglomeration at the census block group scale, by domain. (Unless stated otherwise, variables are proportions expressed in % ; [a] Redundant group "labor force" ; [b] Redundant group "unemployment" ; [c] Not a proportion)

Domain	*Variable name*	*Description*
Education	`AttendingSchool`	People 6-15 years old attending school in the 6-15 years old population
	`NoDiplomas`	People with no diploma (and not studying) in the 15 years old and more population
	`BasicGeneralQualifications`	People with basic or intermediate general or vocation qualifications (and not studying) in the 15 years old and more population
	`GeneralCertificates`	People with general or vocational maturity certificates (and not studying) in the 15 years old and more population
	`LowerTertiaryEducation`	People with at least a lower tertiary education (and not studying) in the 15 years old and more population
	`HigherEducationalDegree`	People with a higher educational degree (and not studying) in the 15 years old and more population
	`Students`	Students in the 15 years old and more population
Housing	`IndividualHouse`	Individual houses in the main residences
	`MultipleDwellingUnits`	Multiple dwelling units in the main residences
	`NonOwner`	Non-owner-occupied in the main residences
	`SubsidizedHousing`	Subsidized housing in the main residences
	`BuiltBefore1968`	Main residences built before 1968
	`Builtafter1990`	Main residences built after 1990
	`Less40m2`	Main residences less than $40m^2$
	`Larger150m2`	Main residences larger than $150m^2$
	`ParkingSpace`	Main residences with a parking space (garage or other)
	`WithoutCar`	Households without a car
	`TwoOrMoreCars`	Households with 2 or more cars

Table 2: Description of the 37 socioeconomic variables available for the Lille agglomeration at the census block group scale, by domain (continued). (Unless stated otherwise, variables are proportions expressed in %).

3. The SesIndexCreatoR package

The **SesIndexCreatoR** package depends on the **FactoMineR** (Husson *et al.* 2013; Lê *et al.* 2008) and **class** (Venables and Ripley 2002) packages. In particular, most of data analysis and visualization functions, such as principal component analysis or hierarchical clustering, used in this package come from **FactoMineR**. We thus refer the user to the **FactoMineR** package and its manual for details on PCA and HC functions outputs. The sources and binaries of the package **SesIndexCreatoR** are available on the Equit'Area website or on CRAN and the installation is standard.

Because the package is also aimed to be used by R novice users, the example data are not included as R dataset but as a text file, in order to show in the package's manual how to import a file.

Function	Description
`ClassifHC`	Internal function: Classification with Hierarchical Clustering (HC)
`ClassifInt`	Internal function: Classification by intervals
`ClassifQuant`	Internal function: Classification by quantiles
`plot.SesClassif`	Plot the results of the classification of a socioeconomic index
`plot.SesIndex`	Plot the results of the construction of a socioeconomic index
`print.SesClassif`	Print the classification of a socioeconomic index results
`print.SesIndex`	Print the creation of a socioeconomic index results
`SelectVar`	Internal function: Selection of variables
`SesClassif`	Create categories from a socioeconomic index
`SesIndex`	Creation of a Socio-Economic Index
`SesReport`	Creation of a report for SesIndex and SesClassif functions
`SesStep1`	Internal function: perform the first step of the creation of the socioeconomic index

Table 3: Functions available in **SesIndexCreatoR** 1.0-1.

SesIndexCreatoR is composed of three main functions and several visualizing and internal functions (see Table 3):

- The `SesIndex` function creates a socioeconomic index such as defined in the Equit'Area project. It is possible to choose the starting set of variables, the potential redundant groups of variables, the potential supplementary units, the method of selection (PCA or MFA) and the step of the procedure to perform. Results include the final index and all the results of the intermediate steps.
- The `SesClassif` function creates socioeconomic categories, based on a socioeconomic index created by `SesIndex` function, with different technics such as hierarchical clustering, quantiles or equals subdivisions. Results include both a table with the original data set with class of each unit and the results of the classification technic (cut points, classes particularities, ...).

- The `SesReport` function creates a .html file with a report summarizing the results of the different steps of the creation of a socioeconomic index with the `SesIndex` function and, if any, the classification of the index using the `SesClassif` function. This function also allows to create a .csv file containing the original data set and the index and, if any, the classification.

4. Example

First, the socioeconomic data from the text file are imported in a data frame:

```
R> library("SesIndexCreatoR")
R> SesData <- read.table(
+    system.file("extdata","SesData.txt", package = "SesIndexCreatoR"),
+    header=TRUE,sep="\t", row.names=1)
```

The SesData.txt contains 37 socioeconomic variables and 1 type variable (giving the type of BG) for each BG of the Lille municipality and adjacent municipalities, as describe in Section 2.1. Then, the SesData dataframe has 234 rows representing the BGs and 38 columns representing the variables.

As the `SesIndex` function needs vectors or lists of variables' names as arguments, we then extract the different vectors and lists needed to call the function (with redundant groups). The first line of the following code chunk allows to extract the names of the variables to analyse as a vector. The remaining lines extract the names of the variables in the two groups of redundant variables (see Table 1) and create a list containing the two vectors of names for the groups of redundant variables.

```
R> varnames <- colnames(SesData)[2:ncol(SesData)]
R> group1 <- grep("+Unemployed", colnames(SesData), value=TRUE)
R> group2 <- grep("+LabourForce", colnames(SesData), value=TRUE)
R> groupvarnames <- list(group1, group2)
```

In order to consider activity and miscellaneous BGs as illustrative units, we extract the names of the corresponding rows (in our example, A is for "Activity" and D for "Miscellaneous" types of BGs) :

```
R> illus <- rownames(SesData[SesData[,"Type"] %in% c("A", "D"),])
```

It is "now" possible to create a socioeconomic index described in Materiel and methods using `SesIndex`. Here, we will create a socioeconomic index using all the 3 steps. Two groups of redundant variables are defined in *groupvarnames* and several BGs are set illustrative. By default, all the 3 steps are performed and step 2 uses a PCA.

```
R> index <- SesIndex(SesData, varnames=varnames, groupvarnames=groupvarnames,
+                    sup=illus)
```

```
R> plot(index, choice="ind", label="none")
```

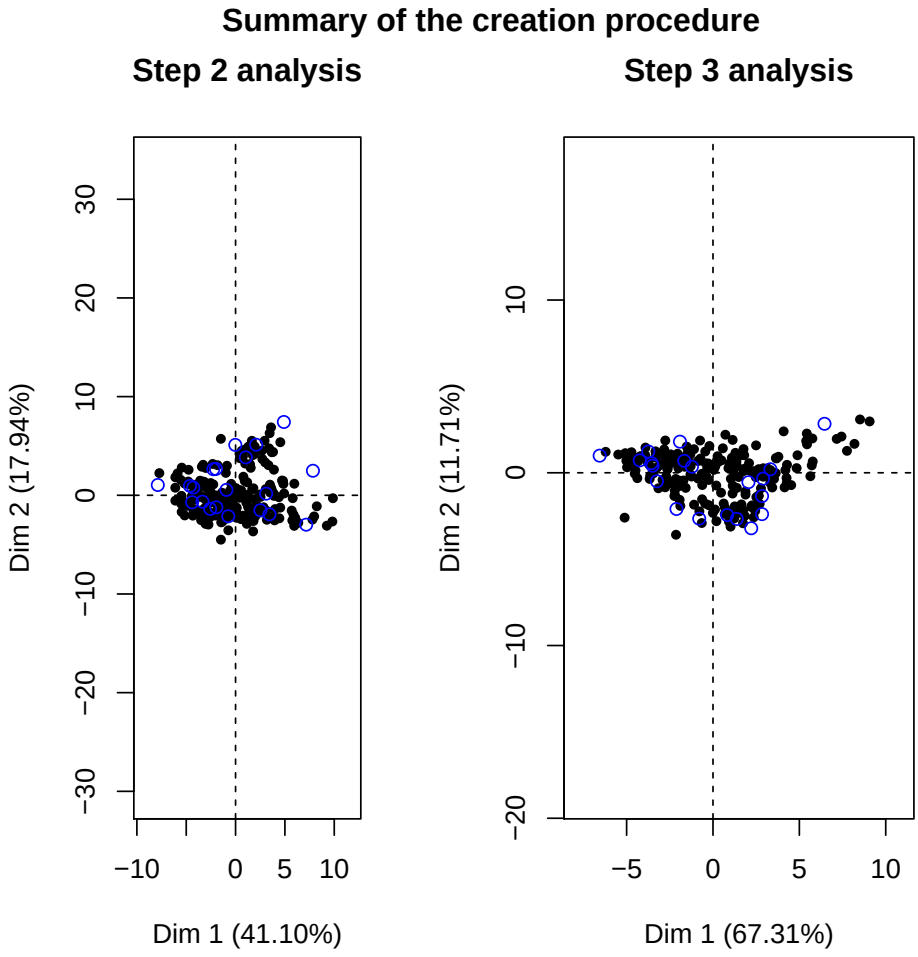


Figure 1: Synthetic view of the graphical outputs for individuals.

Once the index is created, we want to explore the results of the procedure. For instance, among the groups of redundant variables listed in Table 1 (Unemployment and Labor Force), the variables representing the best these groups and selected by our procedure are:

```
R> index$step1$selection
```

```
[1] "UnemployedTotal" "LabourForce"
```

Or, among the list of variables in Tables 1 and 2 (except the redundant variables dropped at step 1), the variables selected to compose the SES index for Lille agglomeration are:

```
R> index$step2$selection
```

```
 [1] "ForeignPop"            "UnemployedTotal"
 [3] "InsecureJobs"          "SteadyJobs"
 [5] "SingleParentFamilies"  "NoDiplomas"
 [7] "IndividualHouse"       "MultipleDwellingUnits"
 [9] "ParkingSpace"          "NonOwner"
[11] "WithoutCar"            "TwoOrMoreCars"
[13] "SubsidizedHousing"     "MedianIncome"
```

```
R> plot(index, choice="var", step=2)
```

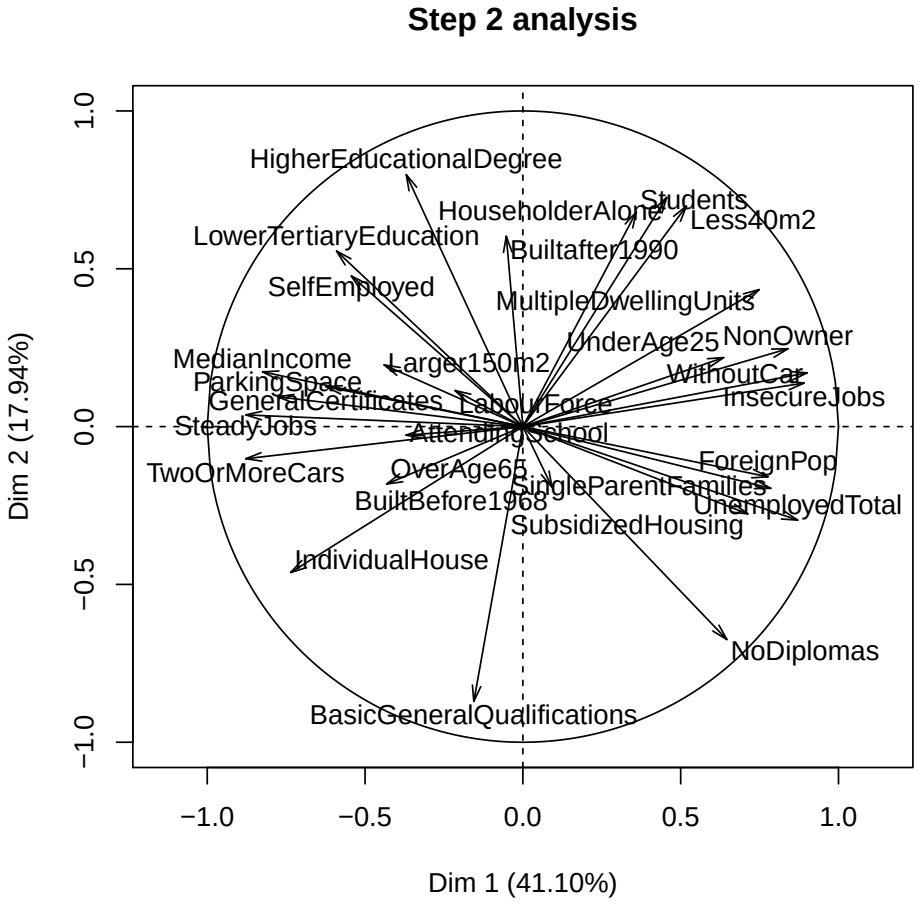


Figure 2: Correlation circle for the second step.

It is also possible to obtain detailed results of the data mining technics, like the correlation coefficients of the variables with the two first components of the second step analysis:

```
R> index$step2$analysis$var$coord[,c(1,2)]
```

```
                                  Dim.1       Dim.2
UnderAge25                   0.63630110  0.21862821
OverAge65                   -0.43163887 -0.18186596
ForeignPop                   0.77678724 -0.15910609
LabourForce                 -0.21536615  0.11435159
UnemployedTotal              0.87073008 -0.29624249
SelfEmployed                -0.54334383  0.47689134
InsecureJobs                 0.89130598  0.13786739
SteadyJobs                  -0.87777733  0.03703831
SingleParentFamilies         0.78556166 -0.19555464
NoDiplomas                   0.64635948 -0.67461860
HouseholderAlone             0.45573846  0.72622501
AttendingSchool             -0.37025379 -0.02709806
BasicGeneralQualifications -0.15542464 -0.87073704
GeneralCertificates         -0.62380039  0.12642215
LowerTertiaryEducation      -0.59037219  0.55626000
HigherEducationalDegree     -0.36942886  0.79823028
Students                     0.35951016  0.68000017
IndividualHouse             -0.73553777 -0.46219873
MultipleDwellingUnits        0.74861629  0.43400686
BuiltBefore1968              0.09240263 -0.19191144
Builtafter1990              -0.05250108  0.60279988
ParkingSpace                -0.77646906  0.09453152
NonOwner                     0.83998699  0.24679532
Less40m2                     0.51827452  0.69853014
Larger150m2                 -0.43976155  0.19532641
WithoutCar                   0.90096827  0.17029037
TwoOrMoreCars               -0.87800029 -0.10183450
SubsidizedHousing            0.71268195 -0.27645939
MedianIncome                -0.82471346  0.17342006
```

Or the proportion of variance explained by the four first components of the final step:

```
R> index$step3$analysis$eig[1:4,]
```

```
       eigenvalue percentage of variance
comp 1  9.4233115              67.309368
comp 2  1.6390996              11.707854
comp 3  1.0678887               7.627776
comp 4  0.5014364               3.581689
       cumulative percentage of variance
comp 1                           67.30937
comp 2                           79.01722
comp 3                           86.64500
comp 4                           90.22669
```

The above outputs are especially interesting to understand the procedure of variable selection. We can see in these results that the variables of total unemployment and total labor force were respectively selected from the groups of redundant unemployment variables and labor force variables. Then, for these two groups only these two variables were kept in the next steps. We can see in the selection from the step 2 that only variables with the highest correlations with the first component were selected. Here, 14 variables out of 29 were kept for the final step and the construction of the SES index. Eventually, the first component of the final step PCA performed on these 14 variables explained more than 67% of the total variance.

```
R> plot(index, choice="var", step=3)
```

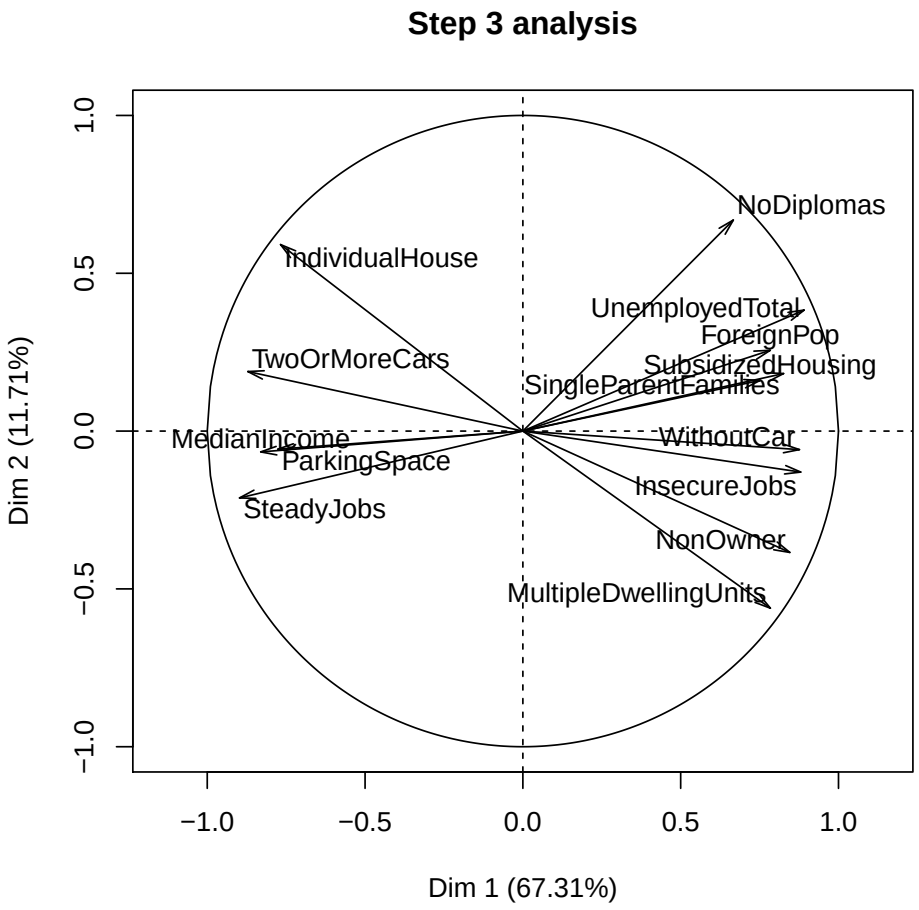


Figure 3: Correlation circle for the final step.

Some graphical outputs can be seen in Figure 1 to Figure 3. Figure 1 is a synthetic view of the projection of the BGs on the first principal components of the PCA performed in step 2

and step 3. Black dots represent active units whereas blue circles represent illustrative units (i.e., BG of the economical activity or miscellaneous types). Due to the number of units, BG labels are not displayed here but are activated by default. The step 3 part of the figure allows to see that BG are mainly along the first component and have not an extremely important variability along the second component.

Figure 2 gives the circle of correlations of the PCA performed in step 2. Most of the variables seem to have a good correlation with the first component, both positively and negatively, whereas as the correlations with the second component are mainly positive (except for two variables). A few variables (5) are not well represented on this plane and may have higher correlations with the third or fourth component. On this figure, a first opposition between "variables of deprivation", at the right, and "variables of favor", at the left, can be seen.

Finally, Figure 3 shows the circle of correlations of the PCA performed in step 3. The opposition between the "deprivation" and the "favor" variables is clear, with a high positive correlation between the first component and proportions of non-owner, unemployment, insecure jobs, person without diploma, subsidized housing, ... and a high negative correlation between the first component and proportion of steady jobs, individual houses, ... The first component of this PCA can then be interpreted as a SES component and be used as a SES index.

We now want to create categories from the socioeconomic index. We use a hierarchical clustering followed by a k-nearest neighbor (k-nn) algorithm. We decide to have an automatic number of classes (i.e., to cut the hierarchical clustering tree where the relative loss of inertia is the highest) :

```
R> categories <- SesClassif(index)
```

Others possibilities currently in the `SesClassif` function are to create classes with hierarchical clustering without k-nn consolidation, with quantiles or with equal range of values.

We can summarize some characteristics of the different categories using simple functions. For instance, it is possible to compare variables average values in eahc category and the overall mean :

```
R> for (i in 1:3) {
+    print(paste("Category",i))
+    print(round(categories$analysis$desc.var$quanti[[i]][,c(2,3,6)],2))
+  }
```

```
[1] "Category 1"
                    Mean in category Overall mean p.value
TwoOrMoreCars                   0.33         0.21       0
IndividualHouse                 0.71         0.45       0
SteadyJobs                      0.73         0.65       0
ParkingSpace                    0.60         0.43       0
MedianIncome                27529.06     21986.21       0
NoDiplomas                      0.11         0.16       0
ForeignPop                      0.02         0.05       0
```

```
SubsidizedHousing                0.08         0.26       0
UnemployedTotal                  0.10         0.16       0
SingleParentFamilies             0.11         0.17       0
MultipleDwellingUnits            0.25         0.52       0
WithoutCar                       0.16         0.29       0
InsecureJobs                     0.09         0.13       0
NonOwner                         0.35         0.58       0
[1] "Category 2"
                     Mean in category Overall mean p.value
MultipleDwellingUnits             0.66         0.52    0.00
NonOwner                          0.69         0.58    0.00
WithoutCar                        0.35         0.29    0.00
InsecureJobs                      0.14         0.13    0.00
SingleParentFamilies              0.18         0.17    0.02
SteadyJobs                        0.63         0.65    0.01
MedianIncome                  19693.52     21986.21    0.00
ParkingSpace                      0.35         0.43    0.00
IndividualHouse                   0.30         0.45    0.00
TwoOrMoreCars                     0.14         0.21    0.00
[1] "Category 3"
                     Mean in category Overall mean p.value
UnemployedTotal                   0.33         0.16       0
ForeignPop                        0.14         0.05       0
SingleParentFamilies              0.28         0.17       0
SubsidizedHousing                 0.74         0.26       0
NoDiplomas                        0.30         0.16       0
InsecureJobs                      0.18         0.13       0
WithoutCar                        0.47         0.29       0
NonOwner                          0.90         0.58       0
MultipleDwellingUnits             0.85         0.52       0
IndividualHouse                   0.13         0.45       0
TwoOrMoreCars                     0.08         0.21       0
ParkingSpace                      0.19         0.43       0
MedianIncome                  12624.39     21986.21       0
SteadyJobs                        0.46         0.65       0
NULL
```

```
R> plot(categories$analysis, choice="map", label="none", draw.tree=F)
```

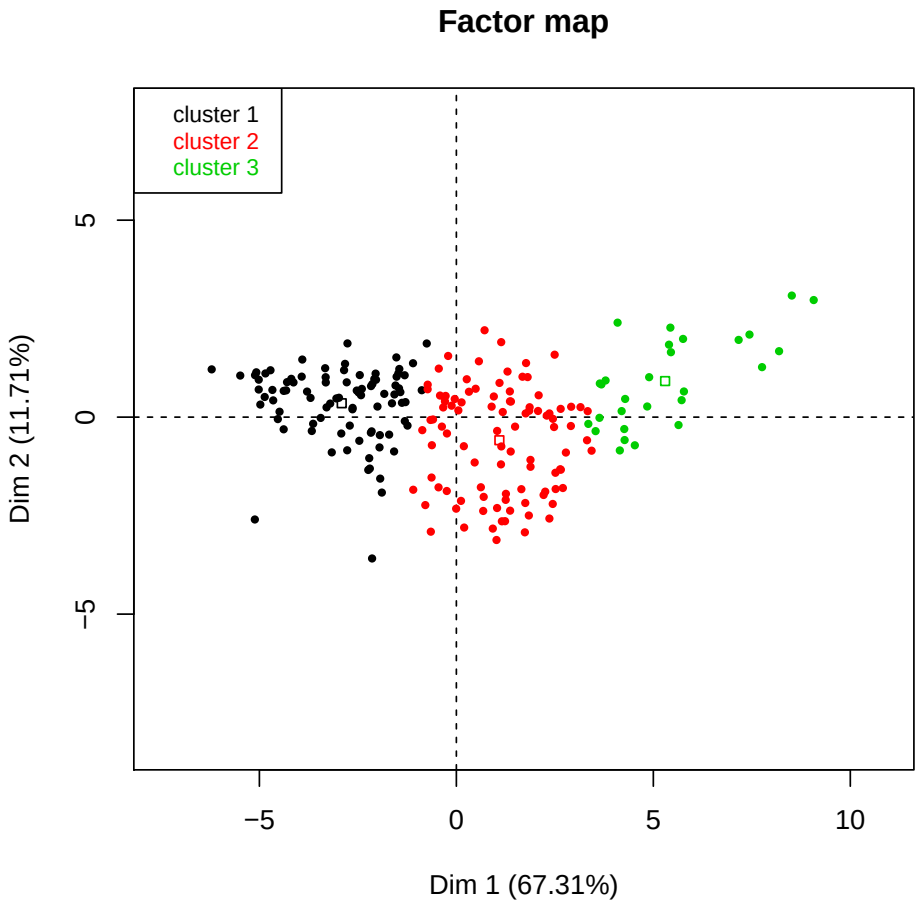


Figure 4: Plot of the individuals by categories.

We can see that the optimal number of categories (according to the inertia criterion) was 3. The description of these categories showed that they are organised by decreasing socioeconomic status. Indeed, category 1 has higher average values of variables like median income or proportion of steady jobs, and lower average values of proportion of unemployed people or proportion of subsidized housing; whereas category 3 has lower values of median income and higher values of unemployment. Figure 4 shows the projection of these categories on the two first axes of the final PCA (note that it is also possible to use directly `plot(categories)` to have both the dendrogram and the projection of the units).

Eventually, we want to export the detailed results of all the three steps of creation of the SES index and of the classification. We also want to export a data file containing the index

and the categories. To do so, the `SesReport` function is used to create .html report (see Appendix). By default, files are created in the current working directory with basename "SesReport" (which can be change as arguments of the `SesReport` function).

```
R> SesReport(categories)
```

5. Conclusion

In this article we presented the **SesIndexCreatoR** package, designed to easily create socio-economic indices with a reproductible statistical procedure. One originality of this procedure compared to other existing indices lays in selecting the final variables for the index by usage of data mining techniques rather than only information gleaned from a literature review, allowing to discard part of the subjectivity that may influence the choice of the variables. This data driven approach allows the data "speak by themself".

The **SesIndexCreatoR** package allows to apply this procedure in a versatile way, by specifying which steps of the procedure should be runned (for instance only step 2 if the aim is to compare selection of variables between metropolitan areas without create indices, or only step 3 if one wants all the introduced variables to be in the index), adding illustrative units or selecting the method used. Once the index created, several tools are available to visualize, synthetize, explore and export the results in a convenient way for further utilization.

We project to extend the package in the future and among other improvements we foresee to implement others methods of classification, to add more tools to help the interpretation of the results, or to allow other ways of visualization (such as mapping). However, these improvement will be made according to users' returns and needs.

References

Bell N, Schuurman N, Hayes MV (2007). "Using GIS-Based Methods of Multicriteria Analysis to Construct Socio-Economic Deprivation Indices." *International Journal of Health Geographics*, **6**, 17.

Carstairs V (1995). "Deprivation Indices: Their Interpretation and Use in Relation to Health." *Journal of Epidemiology and Community Health*, **49 Suppl 2**, S3–8.

Eibner C, Sturm R (2006). "US-Based Indices of Area-Level Deprivation: Results from HealthCare for Communities." *Social Science & Medicine (1982)*, **62**(2), 348–359.

Fukuda Y, Nakamura K, Takano T (2007). "Higher Mortality in Areas of Lower Socioeconomic Position Measured by a Single Index of Deprivation in Japan." *Public Health*, **121**(3), 163–173.

Husson F, Josse J, Le S, Mazet J (2013). ***FactoMineR**: Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.25, URL http://CRAN.R-project.org/package=FactoMineR.

Jarman B (1983). "Identification of Underprivileged Areas." *British Medical Journal (Clinical research ed.)*, **286**(6379), 1705–1709.

Lalloué B, Monnez JM, Padilla C, Kihal W, Le Meur N, Zmirou-Navier D, Deguen S (2013). "A Statistical Procedure to Create a Neighborhood Socioeconomic Index for Health Inequalities Analysis." *International Journal for Equity in Health*, **12**(1), 21.

Lê S, Josse J, Husson F (2008). "**FactoMineR**: An R Package for Multivariate Analysis." *Journal of statistical software*, **25**(1), 1–18.

Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, Culhane J, Elo I, Burke JG, O'Campo P (2006). "The Development of a Standardized Neighborhood Deprivation Index." *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, **83**(6), 1041–1062.

Morris R, Carstairs V (1991). "Which Deprivation? A Comparison of Selected Deprivation Indexes." *Journal of Public Health Medicine*, **13**(4), 318–326.

Padilla C, Deguen S, Lalloué B, Zmirou-Navier D, Viera V (2013a). "Cluster Analysis of Social and Environment Inequalities of Infant Mortality. A Spatial Study in Small Areas Revealed by Local Disease Mapping in France." *Science of the Total Environment*, **454-455**, 433–441.

Padilla C, Lalloué B, Pies C, Lucas E, Zmirou-Navier D, Deguen S (2013b). "An Ecological Study to Identify Census Blocks Supporting a Higher Burden of Disease: Infant Mortality in the Lille Metropolitan Area, France." *Maternal and Child Health Journal*, pp. 1–9.

Pampalon R, Hamel D, Gamache P, Raymond G (2009). "A Deprivation Index for Health Planning in Canada." *Chronic Diseases in Canada*, **29**(4), 178–191.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Salmond C, Crampton P, Sutton F (1998). "NZDep91: A New Zealand Index of Deprivation." *Australian and New Zealand Journal of Public Health*, **22**(7), 835–837.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

B.3 Data Analysis Technics, a Tool for Cumulative Exposure Assessment

Abstract

Benoît Lalloué, Jean-Marie Monnez, Cindy Padilla, Wahida Kihal, Denis Zmirou-Navier, Séverine Deguen

Background: Everyone is subject to environmental exposures from various sources, with negative health impacts (air, water and soil contamination, noise ...) or with positive effects (e.g. green space). Studies considering such complex environmental settings in a global manner are rare.

Aims: We propose to use data mining techniques to create a composite exposure index with a data-driven approach, in view to assess the environmental burden experienced by populations. We illustrate this approach on a large French metropolitan area.

Methods: The study was carried out in the Great Lyon area (France, 1,260,348 inhabitants in 2007, 527 km^2) at the census block group (BG) scale. Indicators on NO_2 annual concentrations, noise levels, proximity to green spaces and to industrial plants, and road traffic were synthetized using Multiple Factor Analysis (MFA), which allows to explore the relations between exposures and BGs without a priori knowledge, and to synthetize indicators of different types. Hierarchical clustering was then used to create BG classes.

Results: The four first components of the MFA explained respectively 30, 16, 12 and 9% of the total variance. Clustering in 5 classes group: 1) A particular type of BGs without population, with large green space; 2) BGs of residential areas, with less traffic, air and noise exposure than average and greener than average; 3) BGs close to industries; 4) BGs of residential areas near midtown; 5) midtown urban BGs, with higher air, traffic and noise exposure than average and less green spaces. Other numbers of classes were tested in order to assess a variety of clustering.

Conclusions: We present a data driven approach using data mining techniques, which seem overlooked for cumulative exposure assessment in complex environmental settings. Although it cannot be applied directly for risk or health effect assessment, the resulting index can help to identify hot spots of cumulative exposure, to prioritize urban policies or to compare the environmental burden across study areas in an epidemiological framework.

Annexe C

Articles d'applications

Dans cette annexe, nous présenterons des informations supplémentaires sur les articles exposant les travaux qui ont été réalisés en collaboration avec les autres membres du projet Equit'Area. On peut présenter ces travaux en deux parties correspondant aux deux volets principaux de la thèse :

- les travaux utilisant le modèle « BYM » détaillé au chapitre 10.2, qui sont présentés en sections C.1 et C.2.
- les travaux utilisant l'indice socio-économique développé au chapitre 7.2, qui sont présentés des sections C.2 à C.7.

On présentera ces articles en suivant à la fois cette distinction suivant les deux volets principaux de la thèse ainsi que par thématiques :

- liens entre la santé et le statut socio-économique (sections C.1 à C.3) ;
- liens entre le statut socio-économique et l'environnement (section C.4) ;
- liens entre la santé, le statut-socio économique et l'environnement (sections C.5 à C.7).

C.1 A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex : a Bayesian modeling approach

L'article intitulé « *A small-area ecologic study of myocardial infarction, neighborhood deprivation, and sex : a Bayesian modeling approach* » a été publié dans la revue *Epidemiology* en juillet 2010. Il s'agit d'un travail préliminaire à la présente thèse qui s'inscrivait dans une étude sur l'influence du statut socio-économique des IRIS et du sexe sur la survenue d'infarctus du myocarde dans la Communauté Urbaine de Strasbourg.

Dans ce travail, le modèle « BYM » est appliqué de manière similaire à celle exposée en section 10.3 pour expliquer les risques d'infarctus du myocarde en prenant comme variable explicative le statut socio-économique contextuel (à l'aide d'un indice socio-économique précédemment développé par l'équipe [137] et ayant servi de base à la procédure développée en section 7.2.1) et en stratifiant les analyses sur le sexe.

Le résultat principal de ce travail est la mise en évidence de risques absolus de survenue d'infarctus du myocarde dans la Communauté Urbaine de Strasbourg plus élevés chez les hommes que chez les femmes. Quel que soit le sexe, un gradient est observé montrant une augmentation du risque relatif des catégories défavorisées par rapport aux catégories favorisées. Ce gradient apparaît linéaire chez les hommes mais pas chez les femmes.

L'une des originalités de ce travail en termes méthodologiques est l'utilisation du modèle « BYM » pour comparer les risques. Une partie importante de cet article est ainsi consacrée à la comparaison entre le modèle « BYM », les modèles utilisant uniquement l'hétérogénéité spatiale non structurée (V_i) ou l'hétérogénéité spatiale structurée (U_i, avec un modèle gaussien conditionnel autorégressif), et le modèle de Poisson sans termes spatiaux.

Abstract

Séverine Deguen, Benoît Lalloué, Denis Bard, Sabrina Havard, Dominique Arveiler and Denis Zmirou-Navier

Background. Socioeconomic inequalities in the risk of coronary heart disease (CHD) are well documented for men and women. CHD incidence is greater for men but its association with socioeconomic status is usually found to be stronger among women. We explored the sex-specific association between neighborhood deprivation level and the risk of myocardial infarction (MI) at a small-area scale.

Methods. We studied 1193 myocardial infarction events in people aged 35?74 years in the Strasbourg metropolitan area, France (2000 ?2003). We used a deprivation index to assess the neighborhood deprivation level. To take into account spatial dependence and the variability of MI rates due to the small number of events, we used a hierarchical Bayesian modeling approach. We fitted hierarchical Bayesian models to estimate sex-specific relative and absolute MI risks across deprivation categories. We tested departure

from additive joint effects of deprivation and sex.

Results. The risk of MI increased with the deprivation level for both sexes, but was higher for men for all deprivation classes. Relative rates increased along the deprivation scale more steadily for women and followed a different pattern: linear for men and nonlinear for women. Our data provide evidence of effect modification, with departure from an additive joint effect of deprivation and sex.

Conclusion. We document sex differences in the socioeconomic gradient of MI risk in Strasbourg. Women appear more susceptible at levels of extreme deprivation; this result is not a chance finding, given the large difference in event rates between men and women.

C.2 An Ecological Study to Identify Census Blocks Supporting a Higher Burden of Disease : Infant Mortality in the Lille Metropolitan Area, France

L'article intitulé « *An Ecological Study to Identify Census Blocks Supporting a Higher Burden of Disease : Infant Mortality in the Lille Metropolitan Area, France* » a été publié dans la revue *Maternal and Child Health Journal* en mars 2013 et s'inscrit dans le travail de thèse que soutiendra Cindy Padilla le 24 octobre 2013 à l'université de Lorraine.

Dans la mesure où ce travail fait également partie du projet Equit'Area, la période d'étude, les données utilisées et l'échelle géographique sont les mêmes que celles présentées dans ce manuscrit. On y utilise à la fois le modèle « BYM » et l'indice socio-économique exposé en section 7.2 pour explorer les liens entre la mortalité infantile et le statut socio-économique, mesurés à l'échelle de l'IRIS dans l'agglomération de Lille Métropole.

Les analyses bayésiennes utilisées dans cet article sont similaires à celles exposées section 10.3. Ainsi, on compare le modèle « BYM » et le modèle de Poisson sans termes spatiaux, avec différentes variables explicatives :

- sans variables explicatives
- avec l'indice socio-économique quantitatif
- avec l'indice socio-économique discrétisé en 5 classes

Le DIC est alors utilisé pour comparer ces modèles et c'est le modèle « BYM » ayant l'indice socio-économique quantitatif comme variable explicative qui apparaît comme ayant la meilleure qualité d'ajustement.

On montre dans cet article une augmentation du risque de mortalité infantile dans les IRIS les plus défavorisés par rapport aux plus favorisés, que l'indice socio-économique soit utilisé comme une variable explicative quantitative ou discrétisé et utilisé comme une variable explicative qualitative. On montre également que les IRIS les plus défavorisés et ayant les plus importants risques relatifs (estimés) se concentrent essentiellement dans les communes de Lille, Tourcoing et Roubaix.

Abstract

Cindy Padilla, Benoit Lalloué, Cheri Pies, Emminarie Lucas, Denis Zmirou-Navier, Séverine Deguen

Background. In France, reducing social health inequalities has become an explicit goal of health policies over the past few years, one of its objectives is specifically the reduction of the perinatal mortality rate. This study investigates the association between infant mortality and social deprivation categories at a small area level in the Lille metropolitan area, in the north of France, to identify census blocks where public authorities should prioritize appropriate preventive actions.

Methods. We used census data to establish a neighbourhood deprivation index whose multiple dimensions encompass socioeconomic characteristics. Infant mortality

data were obtained from the Lille metropolitan area municipalities to estimate a death rate at the census tract level. We used Bayesian hierarchical models in order to reduce the extra variability when computing relative risks (RR) and to assess the associations between infant mortality and deprivation.

Results. Between 2000 and 2009, 668 cases of infant death occurred in the Lille metropolitan area (4.2 per 1,000 live births). The socioeconomic status is associated with infant mortality, with a clear gradient of risk from the most privileged census blocks to the most deprived ones (RR = 2.62, 95% confidence interval [1.87; 3.70]). The latter have 24.6% of families who were single parents and 29.9% of unemployed people in the labor force versus 8.5% and 7.7% in the former.

Conclusion. Our study reveals sociospatial disparities in infant mortality in the Lille metropolitan area and highlights the census blocks most affected by the inequalities. Fine spatial analysis may help inform the design of preventive policies tailored to the characteristics of the local communities.

C.3 Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France

L'article intitulé « *Cluster analysis of social and environment inequalities of infant mortality. A spatial study in small areas revealed by local disease mapping in France* » a été publié dans la revue *Science of The Total Environment* en juin 2013 par Cindy Padilla.

Ce travail fait également partie du projet Equit'Area et en utilise aussi les données. Il étudie les liens entre la mortalité infantile et les variables explicatives de statut socio-économique et de dioxyde d'azote, indicateur de qualité de l'air, sur les agglomérations de Lille et Lyon. Le statut socio-économique est à nouveau représenté par l'indice présenté en section 7.2. L'analyse est par contre effectuée à l'aide d'une autre méthode d'analyse : les modèles additifs généralisés.

Les modèles additifs généralisés ont été introduits par Hastie et Tibshirani au début des années 1990 mêlant les modèles linéaires généralisés et les régressions non-linéaires. Ils sont de la forme $g(\mathbb{E}[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m)$ où l'on spécifie une distribution (de la famille exponentielle) pour Y et une fonction de lien g. Les f_i sont des fonctions paramétriques, semi-paramétriques ou non-paramétriques qui seront alors estimées différemment.

Ici, on ajoute au modèle de Poisson une fonction des coordonnées spatiales (semi-paramétrique) de lissage, ce qui permet de prendre en compte l'aspect spatial des données.

Ce travail met en évidence les motifs de pollution différents entre les agglomérations, mais aussi des relations différentes entre les variables explicatives et la mortalité infantile suivant les villes.

Ainsi, on montre dans le cas de Lille Métropole une forte influence du statut socio-économique sur les variations spatiales de mortalité infantile mais l'on ne retrouve pas cette influence dans le cas du Grand Lyon, même si dans les deux cas les IRIS plus défavorisés ont un taux de mortalité infantile plus important que les IRIS plus favorisés. Dans les deux agglomérations, l'exposition au dioxyde d'azote seul n'explique pas en totalité la distribution spatiale de la mortalité infantile. Enfin, lorsque le statut socio-économique et le dioxyde d'azote sont tous les deux inclus comme variables explicatives, on ne relève plus de variations spatiales du risque de mortalité infantile à Lille mais il demeure une zone ayant un risque plus important dans le cas de Lyon.

Abstract

Cindy Padilla, Séverine Deguen, Benoit Lalloué, Olivier Blanchard, Charles Beaugard, Florence Troude, Denis Zmirou-Navier, Verónica M. Vieira

Background. Mapping spatial distributions of disease occurrence can serve as a useful tool for identifying exposures of public health concern. Infant mortality is an important indicator of the health status of a population. Recent literature suggests

that neighborhood deprivation status can modify the effect of air pollution on preterm delivery, a known risk factor for infant mortality. We investigated the effect of neighborhood social deprivation on the association between exposure to ambient air NO_2 and infant mortality in the Lille and Lyon metropolitan areas, north and center of France, respectively, between 2002 and 2009.

Methods. We conducted an ecological study using a neighborhood deprivation index estimated at the French census block from the 2006 census data. Infant mortality data were collected from local councils and geocoded using the address of residence. We generated maps using generalized additive models, smoothing on longitude and latitude while adjusting for covariates. We used permutation tests to examine the overall importance of location in the model and identify areas of increased and decreased risk.

Results. The average death rate was 4.2‰ and 4.6‰ live births for the Lille and Lyon metropolitan areas during the period. We found evidence of statistically significant precise clusters of elevated infant mortality for Lille and an east-west gradient of infant mortality risk for Lyon. Exposure to NO2 did not explain the spatial relationship. The Lille MA, socioeconomic deprivation index explained the spatial variation observed.

Conclusion. These techniques provide evidence of clusters of significantly elevated infant mortality risk in relation with the neighborhood socioeconomic status. This method could be used for public policy management to determine priority areas for interventions. Moreover, taking into account the relationship between social and environmental exposure may help identify areas with cumulative inequalities.

C.4 Air quality and Social deprivation in four French metropolitan areas ? A spatio-temporal environmental inequality analysis conducted at a small geographical level

L'article intitulé « *Air quality and Social deprivation in four French metropolitan areas ? A spatio-temporal environmental inequality analysis conducted at a small geographical level* » est actuellement en révision dans la revue *Environmental Research Journal*.

Toujours dans le cadre d'Equit'Area, cet article s'inscrit dans un contexte de « justice environnementale », c'est à dire dans l'étude des relations entre les expositions environnementales et le statut socio-économique. Plus précisément, ce travail s'intéresse aux relations spatiales entre l'exposition au NO_2 (telle que présentée en section 6.1) et le statut socio-économique (représenté par l'indice socio-économique) pour les quatre agglomérations du projet Equit'Area, et à l'évolution de ces relations entre deux périodes au cours des années 2000.

Les modèles additifs généralisés sont à nouveaux utilisés, dans ce cas pour chercher à expliquer la moyenne annuelle de NO_2 par l'indice socio-économique ou d'autres variables socio-économiques (proportion d'immigrés, de non propriétaires de leur logement, d'emplois instables, de revenus faibles, ...) en prenant également en compte les effets spatiaux. Chaque agglomération est étudiée indépendamment pour les deux périodes temporelles.

On montre dans cet article que les catégories sociales les plus exposées à la pollution de l'air ne sont pas les mêmes suivant les agglomérations. Ainsi, à Paris ce sont les IRIS les plus favorisés qui sont en moyenne plus exposés au NO_2, tandis qu'à Lille ou Lyon ce sont plutôt les IRIS de statut socio-économique moyen.

On montre également des différences entre les périodes suivant les agglomérations.

Abstract

Cindy Padilla, Wahida Kihal, Verónica Vieira, Benoît Lalloué, Philippe Rosselo, Géraldine LeNir, Denis Zmirou-Navier, Séverine Deguen

Background. Several studies documented that more deprived populations tend to live in areas characterized by higher levels of environmental pollution. Yet, time and geographical patterns of this disproportionate distribution of environmental burden remain poorly assessed, especially in Europe. We investigated the spatial and temporal relation between ambient air NO_2 concentrations and socioeconomic and demographic data in four French Metropolitan Areas (Lille in north of France, Lyon in the center, Marseille in the south and Paris) in two different time periods.

Methods. The geographical unit used was the census block. The response variable was the NO_2 yearly average per census block and the explanatory variables were a neighborhood deprivation index and socioeconomic and demographic data derived from the national census. Generalized additive models allowed to take into account spatial

autocorrelation and generate maps using smoothing on longitude and latitude while adjusting for covariates.

Results. We found that strength and direction of the association between deprivation and NO_2 estimates varied between cities. In Paris, the higher social category seems exposed to a higher concentration of NO_2, on average, whereas in Lyon and in Lille, the census blocks more likely to have higher concentration were those hosting the middle and the lower social categories. Socioeconomic profiles associated with air pollution vary according to the cities and the time period. For Lyon, Marseille and Paris, the proportion of immigrants is so for the first period but not for the second. Conversely, in Marseille and Lille, the proportion of insecure job became linked with NO_2 concentrations in the second period. In Paris the proportion of no owners and low income was so over the two periods.

Conclusion. There is clear evidence of a variety of spatial and temporal patterns of environmental inequalities in French metropolitan areas, entailed to their historical social and built make-up. General statements about environmental inequalities are inappropriate.

C.5 An exploratory spatial analysis to assess the relationship between deprivation, noise and infant mortality

L'article intitulé « *An exploratory spatial analysis to assess the relationship between deprivation, noise and infant mortality* » est en deuxième révision dans la revue *Environmental Health*, il fait partie du post-doctorat de Wahida Kihal.

Avec les mêmes données que celles présentées précédemment, ce travail porte sur les liens entre mortalité infantile et bruit en fonction du statut socio-économique dans l'agglomération de Lyon. Les indicateurs de bruits utilisés sont ceux présentés en section 6.2, et plus particulièrement la moyenne énergétique de « phase 2 » $\bar{L}_{energ,pop}$. Le statut socio-économique est représenté par l'indice précédemment présenté ou différentes variables socio-économiques (revenu, niveau d'éducation, chômage, ...).

La technique utilisée est une méthode d'analyse spatiale appelée « *spatial scan statistics* » fréquemment utilisée dans le contexte d'études spatiales. Cette méthode effectue un balayage de la zone géographique suivant une grille définie automatiquement à partir de la structure spatiale des données ou directement par l'utilisateur (dans le cas présent, les centroïdes des IRIS sont utilisés). Pour chaque point de la grille, différentes fenêtres circulaires autour de ce point sont créées, avec des rayons allant de zéro à une limite supérieure fixée (ici, 50% de la taille de la zone). On compare le taux de mortalité infantile dans chacune de ces fenêtres par rapport au taux attendu sous l'hypothèse d'une distribution aléatoire (on suppose que le nombre de cas dans un IRIS suit une loi de Poisson), et on identifie les groupes les plus vraisemblables d'excès de risque à l'aide d'un test du rapport de vraisemblance.

Il est également possible d'inclure des variables explicatives, dans ce cas le bruit, le statut socio-économique ou les deux (avec interaction).

On montre dans cet article qu'il existe un groupe d'IRIS avec un risque plus élevé de mortalité infantile au sud-est de l'agglomération de Lyon. Ce groupe disparaît (ou se réduit) lorsque l'on ajoute le bruit et le statut socio-économique comme variables explicatives, ce qui suggère que le bruit et le statut socio-économique expliquent l'excès de risque.

Abstract

Wahida Kihal, Cindy Padilla, Benoît Lalloué, Christophe Rougier, Jerôme Defrance, Denis Zmirou-Navier, Séverine Deguen

Background and objective. Few studies explored how noise could contribute to social health inequalities, and even less considered infant mortality or its risk factors as the health event of interest. We investigate in this paper the impact of neighbourhood characteristics, both the socioeconomic status and ambient noise levels, on the spatial distribution of infant mortality in the Lyon metropolitan area, France.

Methods. All infant deaths cases (n=715) that occurred between 2000 and 2009 were geocoded at the census block level. Each census block was assigned socioeconomic

characteristics and Lden levels, a measure of exposure to neighborhood noise. Using a spatial-scan statistic, we examine whether there are significant clusters of high risk of infant mortality according to these neighborhood characteristics.

Results and discussion. Our results highlight that infant mortality is not randomly spatially distributed, with clusters of high risk in the southeast of the Lyon metropolitan area (RR=1.44 ; p= 0.09). After adjustment on socioeconomic characteristics and noise levels, this cluster disappears or shifts according to different scenarios, suggesting that noise exposure and socioeconomic characteristics explain part of the spatial distribution of infant mortality.

Conclusion. Our findings show an impact of noise on the spatial distribution of mortality after adjusting on socio-economic characteristics. Few studies explored the relationship between neighborhood noise and infant mortality which makes difficult the comparison of our finding with others. These results seem plausible in view of 3 hypotheticals not exclusive pathways developed in our conceptual framework: (i) a psychological pathway, (ii) a physiological disruption process and (iii) unhealthy behaviors pathway. These findings require further research for confirmation and interpretation.

C.6 Green space, social inequalities and neonatal mortality in France

L'article intitulé « *Green space, social inequalities and neonatal mortality in France* » est en deuxième révision dans la revue *BMC Pediatrics and child births*.

Toujours dans le cadre du projet Equit'Area et du post-doctorat de Wahida Kihal, ce travail étudie les liens entre la mortalité infantile, les espaces verts et l'influence du statut socio-économique dans le cas du Grand Lyon. Les données sont à nouveau celles déjà présentées, l'indicateur d'espaces verts utilisé étant la proportion d'espaces verts par rapport à la superficie totale de l'IRIS (voir section 6.3.3). Le statut socio-économique est représenté à nouveau soit par l'indice socio-économique, soit par des variables socio-économiques de revenu, de chômage ou d'éducation.

La méthode utilisée est à nouveau le « *spatial scan statistics* », cette fois avec les espaces verts comme variable d'exposition environnementale.

Ce travail est, à notre connaissance, le premier étudiant les liens entre la mortalité infantile, les espaces verts et le statut socio-économique. On y montre à nouveau un groupe d'IRIS avec un risque plus élevé de mortalité infantile dans le sud-est de Lyon mais, là encore, ce groupe disparaît lorsque l'on prend en compte les espaces verts et le statut socio-économique. Ceci tend à suggérer que ces variables expliquent donc l'excès de risque observé dans cette zone.

Abstract

Wahida Kihal, Cindy Padilla, Benoît Lalloué, Marcello Gelormini, Denis Zmirou-Navier, Séverine Deguen

Background: Few studies have considered using environmental amenities to explain social health inequalities. Nevertheless, Green spaces that promote good health may have an effect on socioeconomic health inequalities. In developed countries, there is considerable evidence that green spaces have a beneficial effect on the health of urban populations and recent studies suggest they can have a positive effect on pregnancy outcomes.

Objective: To investigate the relationship between green spaces and the spatial distribution of infant mortality taking account neighborhood deprivation levels.

Methods: The study took place in Lyon metropolitan area, France. All infant deaths that occurred between 2000 and 2009 were geocoded at census block level. Each census block was assigned greenness and socioeconomic deprivation levels. The spatial-scan statistic was used to identify high risk cluster of infant mortality according to these neighborhood characteristics.

RESULTS: The spatial distribution of infant mortality was not random with a high risk cluster in the south east of the Lyon metropolitan area ($p<0.003$). This cluster disappeared ($p=0.12$) after adjustment for greenness level and socioeconomic deprivation, suggesting that these factors explain part of the spatial distribution of infant mortality. These results are discussed using a conceptual framework with 3 hypothetical pathways by which green spaces may have a beneficial effect on adverse pregnancy outcomes: (i) a psychological pathway, (ii) a physiological disruption process and (iii) an environmental pathway.

CONCLUSIONS: These results add some evidence to the hypothesis that there is a relationship between access to green spaces and pregnancy outcomes but further research is required to confirm this.

C.7 Do neighborhood characteristics modify the relation between short-term exposure to nitrogen dioxide and all-cause mortality ? A time-stratified case-crossover study conducted in Paris

L'article intitulé « *Do neighborhood characteristics modify the relation between short-term exposure to nitrogen dioxide and all-cause mortality ? A time-stratified case-crossover study conducted in Paris* » est soumis dans la revue *BMC Public Health* et fait suite au travail de post-doctorat de Claire Petit.

Dans un contexte légèrement différent de celui d'Equit'Area, ce travail vise à étudier l'impact du statut socio-économique et de l'exposition à long terme au NO_2 (à l'échelle du mois ou de l'année) sur la relation entre l'exposition à court terme au NO_2 (à l'échelle de quelques jours) et la mortalité toute-cause des plus de 35 ans à Paris intra-muros.

En effet, si le fait que des variations à court terme de la pollution (autrement dit, des pics de pollution) augmentent la mortalité est désormais bien documenté, l'influence de l'exposition à plus long terme et du statut socio-économique sur cette relation est moins connue.

On utilise ici les données présentées en section 6.1 pour les concentrations annuelles de NO_2. Pour l'exposition à court terme, on utilise une classification des IRIS (également présentée précédemment) qui assigne à chaque IRIS la station de mesure ayant le plus proche profil d'exposition. On utilise alors les concentrations journalières mesurées par cette station comme indicateurs des niveaux de NO_2 dans l'IRIS.

Tous les cas de mortalité de résidents de Paris de plus de 35 ans entre janvier 2004 et décembre 2009 sont considérés, et l'indice socio-économique est utilisé pour représenter le statut socio-économique.

Des données journalières sur la température, l'humidité et la pression atmosphérique de Météo-France et des données sur les cas de grippe du réseau Sentinelle sont également utilisées à des fins d'ajustement.

L'association entre les concentrations journalières de NO_2 et la mortalité est étudiée avec un design dit « cas-croisés ». Ce design est semblable à une étude cas-témoin dans laquelle chaque cas est son propre témoin. On compare alors la période où l'événement sanitaire est présent (ici, la date de décès) à plusieurs périodes où le même sujet ne présentait pas l'événement sanitaire (ici, les mêmes jours de la semaine durant une période d'un mois).

Des modèles de séries temporelles (destinés à supprimer les effets de tendance ou de saisonnalité) ont été employés, ainsi que des régressions logistiques conditionnelles, afin d'analyser ces données.

On montre dans cet article une augmentation de la mortalité toutes causes associée à une augmentation de $10\mu g/m^3$ de NO_2 durant les cinq jours précédents, cette association étant plus importante dans les IRIS les plus défavorisés.

On montre également un excès de risque associé à une exposition à long terme plus importante mais uniquement dans les IRIS les plus défavorisés.

Abstract

Claire Petit, Séverine Deguen, Annabelle Lapostolle, Benoît Lalloué, Wahida Kihal, Cindy Padilla, Tarik Benmarhnia, Pierre Chauvin, Denis Zmirou-Navier

Background. While a great number of papers have been published on the short-term effects of air pollution on mortality, few tried to assess whether this association varies according to the neighborhood socioeconomic level and to long-term ambient air concentrations measured at the place of residence. We explored the effect modification of 1) socioeconomic status, 2) long-term NO_2 ambient air concentrations, and 3) both combined, on the association between short-term exposure to NO_2 and all-cause mortality in Paris (France).

Methods. A time-stratified case-crossover analysis was performed to evaluate the effect of short-term NO_2 variations on mortality, based on 81,453 deaths that occurred among subjects aged over 35 years from 2004 to 2009 in the city of Paris. Stratified analyses were carried out by socioeconomic category, tertiles of modeled long-term NO_2 concentration levels, and both combined. The data were estimated at the census block scale (n=992).

Results. We found an increase of the risk of mortality with a 10 $\mu g/m^3$ short-term increase of NO_2 on a 0-5 lag period. A higher risk of mortality was revealed for subjects living in low socioeconomic areas (excess risk [ER] = 3.51%, (95% CI) = [1.64-5.41], p=0.01) in comparison with the high socioeconomic areas. A higher risk was also suggested in both low socioeconomic and chronically polluted areas (ER = 5.15%, 95% CI = [1.72-8.70], p=0.11).

Conclusion. Our results show that people living in census blocks characterized by a low socioeconomic status are more vulnerable to air pollution episodes; they suggest that living in a high chronically polluted areas could also incur a greater risk after short term episodes.

Méthodes d'analyse de données et modèles bayésiens appliqués au contexte des inégalités socio-territoriales de santé et des expositions environnementales

Cette thèse a pour but d'améliorer et d'appliquer les connaissances concernant les techniques d'analyse de données et certains modèles bayésiens dans le domaine de l'étude des inégalités sociales et environnementales de santé. À l'échelle géographique de l'IRIS sur les agglomérations de Paris, Marseille, Lyon et Lille, l'événement sanitaire étudié est la mortalité infantile dont on cherchera à expliquer le risque avec des données socio-économiques issues du recensement de la population et des expositions environnementales comme la pollution de l'air, les niveaux de bruit et la proximité aux industries polluantes, au trafic automobile ou aux espaces verts.

Deux volets principaux composent cette thèse. Le volet analyse de données détaille la mise au point d'une procédure de création d'indices socio-économiques multidimensionnels et la conception d'un package du logiciel R l'implémentant, puis la création d'un indice de multi-expositions environnementales. Dans cette partie, on utilise des techniques d'analyse de données pour synthétiser l'information afin de fournir des indicateurs composites utilisables directement par les décideurs publics ou dans le cadre d'études épidémiologiques. Le second volet concerne les modèles bayésiens et explique le modèle « BYM ». Celui-ci permet de prendre en compte les aspects spatiaux des données et est mis en œuvre pour estimer le risque de mortalité infantile.

Dans les deux cas, les méthodes sont présentées et différents résultats de leur utilisation dans le contexte ci-dessus exposés. On montre notamment l'intérêt de la procédure de création d'indices socio-économiques et de multi-expositions, ainsi que l'existence d'inégalités sociales de mortalité infantile dans les agglomérations étudiées.

Mots clés : analyse de données, modèles bayésiens, inégalités sociales de santé, expositions environnementales

Data analysis technics and bayesian models applied to the contexte of social health inequalities and environmental exposures

The purpose of this thesis is to improve the knowledge about and apply data mining techniques and some Bayesian model in the field of social and environmental health inequalities. On the neighborhood scale on the Paris, Marseille, Lyon and Lille metropolitan areas, the health event studied is infant mortality. We try to explain its risk with socio-economic data retreived from the national census and environmental exposures such as air pollution, noise, proximity to traffic, green spaces and industries.

The thesis is composed of two parts. The data mining part details the development of a procedure of creation of multidimensional socio-economic indices and of an R package that implements it, followed by the creation of a cumulative exposure index. In this part, data mining technics are used to synthesize information and provide composite indicators amenable for direct usage by stakeholders or in the framework of epidemiological studies. The second part is about Bayesian models. It explains the "BYM" model. This model allows to take into account the spatial dimension of the data when estimating mortality risks.

In both cases, the methods are exposed and several results of their usage in the above-mentioned context are presented. We also show the value of the socio-economic index procedure, as well as the existence of social inequalities of infant mortality in the studied metropolitan areas.

Keywords: data analysis, Bayesian models, social health inequalities, environmental exposures