Santé environnement

Introduction aux statistiques spatiales et aux systèmes d'information géographique en santé environnement

APPLICATION AUX ÉTUDES ÉCOLOGIQUES

Résultats 2010

INSTITUT DE VEILLE SANITAIRE

Sommaire

Introduction aux statistiques spatiales et aux systèmes d'information géographique en santé environnement

Application aux études écologiques

Résultats 2010

Ce travail a été réalisé par **l'Institut de veille sanitaire (InVS)**.

Rédacteurs
Sarah Goria, Morgane Stempfelet, Perrine de Crouy-Chanel, Département santé environnement (DSE)

Relecteurs
Christophe Declercq, Alain Le Tertre, DSE.

Abréviations

CDC	Centers for Disease Control
CMMP	Comptoirs des minéraux et matières premières
DSE	Département santé environnement
EPHT	Environmental Public Health Tracking
EUROHEIS	European Health and Environment Information System
Insee	Institut national de la statistique et des études économiques
InVS	Institut de veille sanitaire
Iris	Ilôts regroupés pour l'information statistique
RIF	Rapid Inquiry Facility
RIVM	National Institute for Public Health and the Environment
RR	Risque relatif
SAHSU	Small Area Health Statistical Unit
SAVIAH	Small Area Variation In Air pollution and Health
SIG	Système d'information géographique
SMARHAGT	SMall ARea Health Analyses: A Geographic Toolkit
SMR	Ratio de morbidité/mortalité standardisée
UIOM	Usine d'incinération d'ordures ménagères

Résumé

Les variations spatiales des indicateurs de santé et des facteurs d'expositions environnementales sont étudiées en épidémiologie dans un but descriptif et afin de suggérer des hypothèses étiologiques.

Ce travail s'intéresse aux études écologiques dans lesquelles les données (indicateurs de santé et facteurs de risque) sont mesurées à l'échelle d'une unité géographique (commune, îlots regroupés pour l'information statistique (Iris), etc.) et non à l'échelle de l'individu.

D'une part, les études autour d'un point source visent à déterminer s'il existe ou non un excès de risque lié à l'exposition générée par un site particulier. D'autre part, les études de corrélations géographiques ont pour objectif d'examiner d'éventuelles relations entre les variations spatiales de facteurs d'exposition environnementale et celles des indicateurs sanitaires.

L'objectif de ce travail est de présenter et de discuter sur les principaux outils et méthodes mettant en œuvre des systèmes d'information géographiques (SIG) et les statistiques spatiales utilisées dans les études écologiques géographiques. Les possibilités qu'offrent la mise en œuvre des SIG et l'exploitation des données géographiques sont présentées en s'appuyant sur des exemples concrets de travaux menés au Département santé environnement (DSE) de l'Institut de veille sanitaire (InVS), ainsi que quelques exemples issus de la littérature, en insistant sur les précautions qui doivent accompagner leur utilisation.

Sont ensuite décrites et discutées des méthodes statistiques adaptées à l'analyse de données agrégées et à l'analyse des relations entre indicateurs sanitaires et indicateurs d'exposition à des facteurs de risques environnementaux. La modélisation et l'analyse statistique de ces données posent un certain nombre de difficultés méthodologiques : la forte variabilité, la dépendance spatiale, l'existence de différentes échelles spatiales, etc. Sont présentés les outils statistiques les plus utilisés pour répondre à ces difficultés.

Malgré un certain nombre de biais et de difficultés d'interprétation liés précisément à la nature agrégée des données, les études écologiques présentent certains avantages, notamment en termes de puissance statistique, d'étendue de la zone et de la population d'étude. De nombreux travaux sont consacrés au développement méthodologique des études écologiques géographiques en santé-environnement et concernent en particulier les méthodes de détection de clusters, les modèles spatiaux, spatio-temporels, les modèles conjoints de plusieurs maladies ou de sources de données multiples, etc. Les travaux de développement méthodologique tentent de réduire les biais inhérents aux études écologiques. Combiner des données d'exposition individuelles ou intra-unité spatiale avec les données agrégées permettrait d'améliorer ce type d'étude. Concernant les études d'investigations autour d'un point source, il semble plus pertinent de réaliser une étude multicentrique autour de sites présentant les mêmes caractéristiques d'émission.

1. Introduction

"L'épidémiologie spatiale est de plus en plus utilisée pour évaluer des risques sanitaires en association avec des pollutions environnementales. Pour cela, elle doit combiner des méthodes de l'épidémiologie, des statistiques et les techniques des systèmes d'information géographique" [1].

Dans le champ de la santé environnementale, l'analyse de la répartition spatiale d'indicateurs de santé comporte différents objectifs : d'une part, la description de ces variations et la modélisation de leur structure et, d'autre part, la mise en évidence des associations entre ces variations et celles d'expositions à des facteurs de risque environnementaux.

Les variations spatiales des indicateurs de santé et des facteurs d'expositions environnementales sont étudiées en épidémiologie dans un but descriptif et afin de suggérer des hypothèses étiologiques [2].

Différents types d'analyse faisant intervenir une approche géographique peuvent être identifiés : la représentation cartographique du risque de maladie, la détection d'agrégats spatiaux de cas (clusters) autour d'un point source, par exemple (incinérateurs, sites de stockage de déchets radioactifs…), ou encore l'évaluation de l'association entre risque et exposition environnementale en fonction de facteurs de risque connus (cancer et rejets des incinérateurs, cancer et arsenic hydrique, par exemple). Pour effectuer ces analyses, des informations spatialisées sont mobilisées. Ces informations peuvent être des données sanitaires, comme par exemple les cas pour une pathologie donnée, géolocalisées à partir des adresses présentes dans les bases de données existantes ainsi que des informations contextuelles (occupation du sol, réseaux routiers, les sites pollués, etc.) qui seront exploitées dans un SIG puis dans l'analyse statistique.

Ce travail s'intéresse plus particulièrement aux études écologiques dans lesquelles les variables (indicateurs de santé et facteurs de risque) sont mesurées à l'échelon d'un groupe (unité géographique, commune, Iris, etc.) choisi pour sa pertinence, selon l'objectif de l'étude et les données dont on dispose, et non à l'échelle de l'individu. L'indicateur sanitaire est alors une donnée agrégée qui est le résumé d'observations individuelles comme par exemple le nombre observé de cas de cancer par commune. Les études écologiques présentent plusieurs avantages. En effet, elles utilisent des données déjà disponibles, grâce, par exemple, aux registres de maladies (cancers, malformations congénitales, notamment) et au développement des pratiques de géoréférencement des adresses des cas enregistrés. Elles exploitent des estimations de facteurs environnementaux réalisées à l'échelle des unités géographiques, qui présentent *a priori* des variations inter-unités plus importantes qu'entre individus. Elles permettent en outre de se prémunir du risque d'erreurs de mesure lié aux études individuelles. Ces études, qui utilisent comme unité d'observation l'ensemble de la population résidant dans une zone, ont une puissance statistique importante. Cependant, elles posent également un certain nombre de problèmes méthodologiques sur lesquels nous reviendrons. Parmi ces difficultés, celle de l'interprétation des résultats est délicate : l'évaluation du lien individuel à partir du lien estimé sur les données agrégées est souvent loin d'être directe et évidente. Sous la dénomination générale d'étude écologique sont en fait regroupés plusieurs types de travaux aux démarches méthodologiques différentes. Nous nous intéresserons ici plus particulièrement, d'une part, aux études menées autour d'un point source, et, d'autre part, aux études de corrélations géographiques. Les études autour d'un point source visent à déterminer s'il existe ou non un excès de risque lié à l'exposition générée par un site particulier. Les études de corrélations géographiques, également dénommées régressions écologiques, ont pour objectif d'examiner d'éventuelles relations entre les variations de facteurs d'exposition environnementale et les variations d'un ou plusieurs indicateurs sanitaires. Il ne s'agit pas ici de dresser une liste exhaustive des différentes méthodes et outils de l'approche spatiale en épidémiologie. L'objectif de ce travail est de présenter et de discuter sur les principaux outils et méthodes mettant en œuvre des systèmes d'information géographiques et les statistiques spatiales utilisées dans les études écologiques géographiques.

Après avoir précisément défini les spécificités des études écologiques géographiques auxquelles s'intéresse plus spécifiquement ce travail, nous définirons les SIG ainsi que les grandes caractéristiques des données spatialisées. Les possibilités de tels outils et données seront présentées en s'appuyant sur des exemples concrets de travaux menés au DSE de l'InVS, ainsi que quelques exemples issus de la littérature, en insistant sur les précautions qui doivent accompagner leur utilisation. De la définition d'une population ou d'une zone d'étude, à la cartographie des résultats, en passant par l'identification des sources et des voies d'exposition potentielles (industries polluantes, principales infrastructures routières, etc.) ou l'estimation de l'exposition des populations (création d'indicateurs d'exposition, etc.), l'utilisation des SIG trouve tout son sens dans des applications à la santé environnementale tant le volume et la diversité des données à mobiliser peuvent parfois être importants.

Sont ensuite abordées et discutées des méthodes statistiques adaptées à l'analyse de données agrégées – l'agrégation des données sanitaires à l'échelle d'une maille géographique donnée étant la caractéristique première des études

écologiques – et à l'analyse des relations entre indicateurs sanitaires et d'exposition à des facteurs de risque environnementaux. La modélisation et l'analyse statistique de ces données posent un certain nombre de difficultés méthodologiques sur lesquelles nous reviendrons plus longuement : la forte variabilité, la dépendance spatiale, l'existence de différentes échelles spatiales, etc. Nous présenterons les outils statistiques les plus utilisés pour répondre à ces difficultés, qui sont les modèles hiérarchiques bayésiens mis en œuvre grâce aux techniques de Monte Carlo par chaînes de Markov.

Afin d'ouvrir cette présentation à des innovations méthodologiques, le chapitre 5 de ce travail s'intéresse à l'outil Rapid Inquiry Facility (RIF), développé par l'Imperial College de Londres, qui combine méthodes statistiques et technologie SIG pour des études épidémiologiques.

Enfin, ce travail s'achève par des conseils sur l'utilisation de certaines méthodes utilisées en épidémiologie géographique (ou la recommandation de ne pas les utiliser).

Afin de bien délimiter le cadre des types d'études auxquels s'intéresse ce travail, il convient, en premier lieu, de définir précisément les caractéristiques ainsi que les difficultés spécifiques qui sont liées aux études écologiques géographiques.

1.1 Études locales autour d'un point source

Les études descriptives effectuées autour de sites particuliers cherchent à mettre en évidence une augmentation du taux d'incidence de pathologies ou de taux de mortalité spécifiques de populations vivant à proximité de ces sites.

La population vivant à proximité d'une source polluante est supposée exposée et est comparée à une population de référence supposée non exposée ou, en tout cas, moins exposée. Ces études se retrouvent sous différentes formes selon la zone de référence choisie. Il peut s'agir d'une comparaison :
- zone locale *versus* zone de référence (en général, la zone de référence est la France entière) ;
- zone locale au sein de zones voisines de tailles équivalentes ;
- zone locale avec un gradient d'exposition.

L'objectif est de conclure ou ne pas conclure à une différence entre la population exposée et la population de référence et répondre à la question sur la nécessité ou la non-nécessité de faire d'autres études.

Le nombre de cas survenus sur une période donnée est comptabilisé, le nombre de cas attendus est estimé à partir de taux d'incidence/mortalité de référence (taux "France entière", par exemple) et l'existence d'un excès ou d'un déficit de cas est alors testé en comparant le nombre de cas observés au nombre de cas attendus, standardisés sur des facteurs démographiques, l'âge et le sexe, principalement. Ce ratio d'incidence/mortalité standardisé représente le déficit ou l'excès d'incidence/mortalité par rapport à une population type.

L'approche classiquement utilisée pour l'analyse du risque à proximité d'un point source consiste dans le calcul du ratio de morbidité/mortalité standardisée (SMR), de son intervalle de confiance et d'un test statistique.

Le chapitre 2 présente des outils SIG qui peuvent être utiles dans ces études pour définir et caractériser la population exposée, délimiter la zone d'intérêt, décrire un contexte environnemental, etc. Le chapitre 3 présente des méthodes statistiques alternatives qui peuvent être utilisées dans ces études : les méthodes de détections de cluster et de représentation cartographique.

La note méthodologique "Les études locales autour d'un point source. Les différentes méthodes statistiques, leurs avantages et leurs inconvénients" [3] présente les méthodes utilisées, et discute des limites et difficultés de ce type d'étude.

"Ces études posent des problèmes quant à la méthodologie d'analyse ainsi qu'au niveau de l'interprétation des résultats"[4]. Les études locales autour d'un point source souffrent d'être mises en place pour répondre à une perception déjà existante d'une surincidence – en effet, ces études sont souvent menées pour répondre à l'interrogation des populations locales à proximité de ces points sources. Les tests statistiques sont alors utilisés *a posteriori* "pour confirmer" l'éventuelle surincidence. Les hypothèses des tests statistiques, la collecte des données et la définition de la zone d'étude peuvent être biaisés par ce problème [5]. De plus, les études locales souffrent de leur construction intrinsèque basée sur une analyse écologique géographique (biais écologique), d'un manque de contrôle des facteurs de confusion, de leur caractère unique, les rendant potentiellement sujettes aux variations aléatoires. Ces défauts rendent difficilement crédible toute analyse locale basée sur cette approche, sauf dans le cas exceptionnel où le risque serait tel qu'il ne pourrait pas *a priori* être lié à un facteur de confusion ou au hasard.

1.2 Études de corrélations géographiques

L'objectif est d'étudier, au niveau de groupes d'individus définis sur une base géographique, la relation entre un indicateur de santé et une exposition environnementale.

Les études écologiques n'ont pas pour but l'étude des risques au niveau individuel mais l'étude des effets de groupe expliquant une partie de la variation entre les unités géographiques de l'incidence de la pathologie étudiée. Il s'agit d'études descriptives qui peuvent permettre de générer des hypothèses étiologiques individuelles [6]. Les études de corrélations géographiques sont appropriées quand il s'agit d'étudier des expositions stables dans le temps mais variables dans l'espace (radon, composition de l'eau de boisson).

Pour réaliser ces études, il est nécessaire de définir : les pathologies d'intérêt et les indicateurs sanitaires pertinents, l'unité statistique, la zone d'étude et la période d'étude, le facteur de risque environnemental d'intérêt et l'indicateur d'exposition à ce facteur de risque ainsi que les facteurs de confusion.

Les indicateurs sanitaires sont basés en général sur les données recueillies en routine, telles que les données de mortalité issues de la surveillance pérenne des causes médicales de décès ou sur des données d'incidence. Pour les cancers, ces données d'incidence peuvent être issues de la surveillance pérenne des registres des cancers (registres généraux, registres spécifiques). Les données de population par sexe, âge et unité spatiale sont essentielles pour calculer les indicateurs sanitaires. Ces données sont obtenues par recensement et sont disponibles auprès de l'Institut national de la statistique et des études économiques (Insee). Les difficultés liées à l'estimation de la population dans ce type d'étude sont présentées dans une note méthodologique déjà citée [3].

Unité spatiale

Un point clé de ces études est le choix de l'unité géographique appropriée. L'unité optimale doit être assez grande pour fournir des indicateurs de santé stables et assez petite pour être homogène en termes d'exposition, de caractéristiques socio-économiques… Souvent, l'unité est choisie en fonction des données de santé et des données démographiques disponibles et est définie sur un découpage géographique de type administratif (commune, canton) ou statistique (Iris). La nature administrative du découpage peut amener à une très forte hétérogénéité dans la répartition démographique avec des zones peu peuplées (zones rurales) et des zones densément peuplées (zones urbaines). Par ailleurs, ce découpage administratif peut ne pas être toujours pertinent d'un point de vue épidémiologique, et les résultats peuvent être sensibles à sa redéfinition.

Zone d'étude

La zone d'étude doit permettre une bonne discrimination de l'exposition. Elle doit permettre aussi de disposer d'une population suffisamment large pour les données sanitaires.

Période d'étude

La période d'étude repose habituellement sur les données les plus récentes. En fonction des pathologies étudiées et de la latence de leurs survenues par rapport à l'exposition, elle peut refléter une exposition antérieure, allant de plusieurs années à plusieurs dizaines d'années. Tout comme pour la zone d'étude, afin de pallier au manque d'effectif, plusieurs années de données sont collectées. Aucune information individuelle n'étant disponible, les mouvements migratoires ne sont pas pris en compte : on suppose que les individus n'ont pas déménagé entre la période d'exposition et la période d'étude. Ceci rend difficile une bonne définition de la population à risque au niveau de petites zones géographiques.

Facteurs de confusion

Un autre point important est la prise en compte des facteurs de confusion "en raison de la faiblesse quantitative des risques estimés, faiblesse qui rend théoriquement plus plausible qu'une partie de l'effet provienne de variables concomitantes."[7].

Méthodes et biais

Le chapitre 2 présente des outils SIG utiles, voire indispensables, pour mener des études de corrélations géographiques, notamment pour la construction d'indicateurs d'exposition ou de confusion (indicateur de densité de l'habitat par commune par exemple à partir du bâti), pour caractériser un contexte environnemental (occupation du sol, topographie, etc.). Le chapitre 3 présente les méthodes statistiques utilisées dans ces études.

On trouve de nombreux exemples de ce type d'études dans la littérature [8-12]. Les études UIOM et cancer [13] et arsenic hydrique et cancer (rapport en cours d'écriture) sont des exemples de ce type d'étude menées par le DSE de l'InVS.

Ces études sont difficiles à interpréter au niveau individuel à cause du biais écologique, c'est-à-dire la différence potentielle entre le lien dose-effet individuel et celui estimé au niveau de groupe. Le biais écologique est dû à la variabilité intra-unité de l'exposition et des facteurs de confusion. Les conséquences de cette variabilité sont : un biais de spécification (non prise en compte au niveau du groupe de relations dose-effet individuelles non linéaires), un biais de confusion (non-prise en compte de facteurs de confusion) et un biais de standardisation (les indicateurs d'exposition et de santé ne sont pas standardisés sur les mêmes facteurs de confusion – âge, sexe).

Le biais écologique peut être négligeable dans le cas d'une faible variabilité intra-unité de l'exposition, qui peut être obtenue en réduisant la taille des unités spatiales, ou d'un faible lien écologique. Le biais écologique peut être réduit en utilisant de l'information sur la variabilité intra-unité en incorporant des données individuelles dans le modèle, en introduisant les facteurs de confusion potentiels et en incorporant des effets aléatoires dans le modèle. Pour qu'il n'y ait pas de biais de standardisation, si on a standardisé l'indicateur sanitaire sur l'âge, par exemple, il faudrait que l'exposition soit constante entre classes d'âge. Il convient de se référer aux articles de Wakefield [14], de Salway [15] et de Wakefield and Salway [16] pour une revue de ces biais et des méthodes pour les réduire. Best *et al.* [8] ont étudié la sensibilité des analyses de régression écologique à différents biais et à la présence d'erreurs dans les données. De plus, il est difficile de tenir compte d'un temps de latence approprié entre l'exposition et l'effet sur la santé.

Avant d'entreprendre une étude écologique, quelques éléments doivent être considérés avec attention [15] :
- il est important que la variabilité de l'exposition entre unités soit élevée et que la variabilité intra-unité soit faible ;
- il est important de prendre en compte le plus d'informations possible sur les facteurs de confusion potentiels ;
- il est important de prendre en compte l'influence de facteurs de confusion non mesurés avec des modèles adaptés (par exemple, en cas de surdispersion : modèles avec effets aléatoires) ;
- il est important que la variabilité entre unités des facteurs de confusion non mesurés soit la plus petite possible.

Les facteurs qui donnent confiance dans les résultats de ce type d'études sont la qualité des données, l'utilisation appropriée des données et la prise en compte des limites des données. Les limites des données affectent les résultats des analyses statistiques et limitent les analyses qui peuvent être faites. Elles doivent être prises en compte au moment du choix du type d'étude.

1.3 Intérêt et limites de ces études

Les avantages de ces études sont :
- données facilement disponibles (registres des maladies, recensement) ;
- la taille de la population étudiée peut être importante, ce qui facilite la détection d'augmentations de risque relativement faibles ;
- les mesures moyennes au niveau d'unités géographiques sont sans doute moins sujettes à des erreurs de mesure ;
- les contrastes d'exposition sont potentiellement plus importants qu'au niveau individuel (augmentation de la puissance) [8] ;
- elles correspondent à des "expériences naturelles" quand l'exposition a une base géographique physique (radon, pollution de l'air, qualité de l'eau) qui peut être exploitée [8].

Les développements statistiques de l'équipe du Small Area Health Statistical Unit (SAHSU) et l'utilisation des SIG ont ramené de l'intérêt vers les études écologiques en particulier pour des petites unités spatiales.

Ces études doivent être interprétées avec prudence à cause des nombreuses sources de biais et de confusion, en particulier à cause du biais écologique et des erreurs de classification de l'exposition.

Les modèles de régression écologiques font abstraction de l'information sur les expositions individuelles et leur variabilité. Le biais écologique peut être important, par exemple, quand il y a de la variabilité intra-unité de l'exposition

et le modèle exposition-risque n'est pas linéaire, et quand il y a de la variabilité intra-unité des facteurs de confusion [8]. Le biais écologique peut être négligeable dans le cas d'une faible variabilité intra-unité ou d'un faible lien écologique.

Les études écologiques restent attractives pour étudier des facteurs environnementaux pour lesquels l'exposition est relativement homogène et si l'unité géographique utilisée est assez fine. De plus, obtenir des expositions individuelles est difficile. La possibilité que le biais écologique invalide les conclusions d'une étude écologique montre que l'utilité de ce type d'étude est limitée [15]. C'est pour cela qu'avant d'engager une étude écologique, il est nécessaire de vérifier la disponibilité des données et, en particulier, des facteurs de confusion potentiels. Pour prévenir le biais écologique, il serait nécessaire d'avoir des données individuelles pour contrôler la distribution intra-unité de l'exposition et des facteurs de confusion et effectuer alors des études qui combinent données écologiques et données individuelles [17-19]. La prise en compte de la variabilité de l'exposition intra-unité géographique peut améliorer les estimations des effets individuels [20].

2. Les systèmes d'information géographique

L'objet de ce travail étant de présenter les outils et les méthodes mettant en œuvre les SIG et les statistiques spatiales appliquées aux études géographiques en santé environnementale, il est nécessaire pour commencer de définir précisément ce qu'est un SIG.

"Un SIG est un ensemble de matériels informatiques, de logiciels, de données géographiques, et de personnel capable de saisir, stocker, mettre à jour, manipuler, analyser et présenter toutes formes d'informations géographiquement référencées (F. de Blomac, 1994)". Un SIG est de ce fait un outil complet de connaissance, d'aide à la décision et de communication. Il ne peut être réduit à sa seule dimension de logiciel, même si dans le langage courant, c'est assez souvent le cas. Enfin, un SIG ne saurait être réduit à la fonctionnalité de cartographie automatique même s'il permet de produire des cartes qui restent un outil privilégié de réflexion et d'information. Les sciences de l'information géographique ou "géomatique" constituent ainsi un réel savoir-faire que l'on ne saurait limiter à la seule connaissance et maîtrise d'un logiciel informatique.

La plupart des écrits sur le sujet s'attardent justement sur la dimension "outil informatique" de ces systèmes tandis que d'autres insistent sur les capacités d'aide à la décision et de production de nouvelles informations à partir de la superposition de données préalablement spatialisées [21].

Dans tous les cas, le SIG permet la collecte et le stockage, la visualisation, la superposition, l'interrogation et l'analyse des données géoréférencées. Sa logique systémique en constitue le principal atout dans l'identification et l'évaluation des questions de tous ordres. Dans le champ de la santé, il permet l'étude des dynamiques spatiales pour la compréhension de certains faits de santé. Ceci afin de répondre aux questions : qui ? Où ? Et pourquoi là ?

Dans le champ plus spécifique de la santé environnementale, les SIG sont notamment utilisés dans le processus d'évaluation de l'exposition des personnes. Ils permettent de préciser la délimitation spatiale de la population étudiée (à l'aide des données d'occupation du sol, par exemple), d'identifier des sources et des voies d'exposition potentielles (industries polluantes, principales infrastructures routières, etc.), d'intégrer à l'analyse les niveaux de certains polluants dans l'environnement pour finalement estimer l'exposition des populations (création d'indicateurs d'exposition) [22].

2.1 Les possibilités d'utilisations du SIG au regard des besoins en santé environnementale

L'atout principal du SIG réside dans le fait qu'il permet de visualiser, d'explorer et de croiser plusieurs sources de données très différentes en même temps. Ce type d'exploration permet de mieux comprendre d'éventuelles interrelations entre l'environnement, la santé et les caractéristiques démographiques et/ou socio-économiques des populations.

D'une manière générale en santé, les SIG sont aussi beaucoup sollicités pour la représentation cartographique des maladies. Par ailleurs, le SIG rend possible la mise à jour des données pour suivre une évolution spatio-temporelle d'un problème de santé. Enfin, il peut être très utile comme outil d'aide à la décision dans le cadre d'une procédure de gestion des alertes. Il permet la synthèse des données de gestion de crise pour une compréhension rapide des enjeux et une prise de décisions.

En épidémiologie environnementale, plus particulièrement, l'approche spatiale est prépondérante dans plusieurs types d'études : les études écologiques géographiques et les études locales autour de points sources. C'est sur ces deux types d'études que porte plus particulièrement ce travail. Mais le SIG n'en est pas moins également mobilisé dans des études épidémiologiques individuelles dès lors qu'il s'agit d'étudier une exposition. Les SIG sont très souvent associés à ces types d'études lorsqu'elles comportent une forte composante spatiale.

On peut classer les différentes utilisations qui peuvent être faites des SIG en santé environnementale de la façon suivante :

- **la localisation des données sanitaires et environnementales pour leur visualisation** : le SIG permet de géoréférencer (attribuer des coordonnées géographiques à un objet afin de le localiser dans l'espace) des données et de mettre en œuvre des données de nature très différentes et qui n'ont souvent pas la même résolution spatiale. Grâce à la prise de connaissance de ces données, c'est l'ensemble du contexte environnemental qui peut être décrit et mieux appréhendé ;
- **la délimitation de la zone d'étude et la description des populations exposées** : c'est souvent dans le SIG, une fois intégrées les données sanitaires après géo-référencement, et les données du contexte environnemental, que pourra être définie la zone d'étude et de ce fait, la population d'intérêt. Dans le cas d'une étude locale autour d'un point source, le SIG permet d'investiguer rapidement en visualisant précisément le lieu de l'incident (accès, distances, environnement, importance et répartition, la délimitation d'un périmètre d'intervention et/ou de sécurité, etc.), en localisant les dangers recensés sur le territoire, en représentant un contexte environnemental à travers des équipements et des points de vulnérabilité (dans le cas d'un risque de pollution de l'eau par exemple, c'est dans le SIG que l'on pourra rapidement positionner les unités de distribution, les points de captages, etc.), pour finalement aider à la définition de la zone d'étude et à l'évaluation d'une exposition de la population et son niveau de vulnérabilité (positionnement des établissements de santé, des écoles, des crèches, etc.) ;
- **la construction d'indicateurs** : le SIG est utilisé pour combiner les données afin de créer de nouvelles informations (par exemple, des indicateurs d'exposition au trafic, indicateur d'exposition à une pollution de type industriel, etc.) pour permettre l'analyse statistique (détection de cluster, régression de Poisson, etc.) ;
- **la communication** d'une information : enfin, c'est grâce aux outils du SIG que sera cartographiée l'information de manière efficace et directement utilisable pour la prise de décision et la communication des résultats d'une étude épidémiologique.

2.2 Définitions et notions clefs indispensables à la mise en place d'un SIG

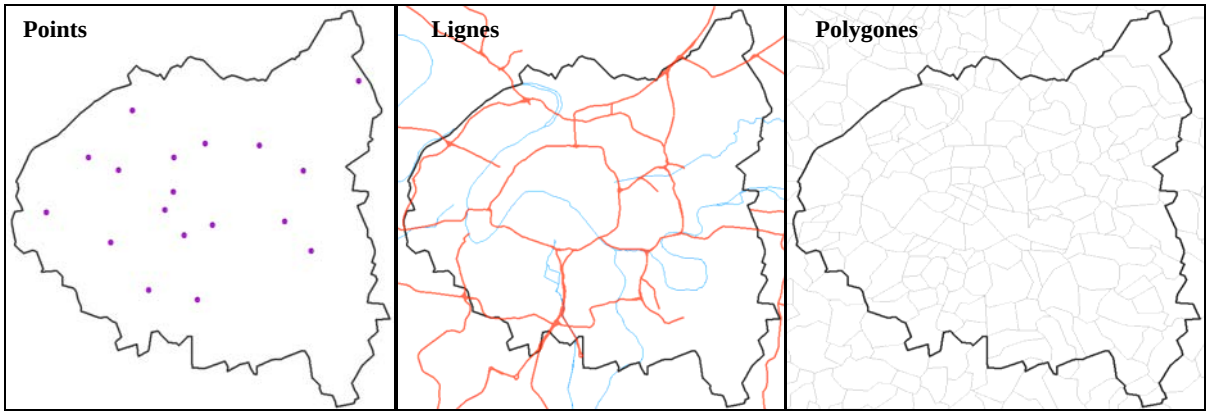
2.2.1 En épidémiologie, qu'est-ce qu'un objet géographique ?

Les objets géographiques rencontrés en épidémiologie ne sont pas différents de ceux rencontrés dans d'autres domaines. Toute représentation cartographique passe par la traduction des éléments réels que l'on observe (industrie, route, limite administrative) en objets graphiques qui sont de trois types (figure 1) :

- le point : par exemple, les lieux de résidence des sujets atteints de la pathologie étudiée (cas) géoréférencés et représentés sur une carte ou encore les sites industriels, les hôpitaux, etc. ;
- la ligne (le linéaire) : les routes, les cours d'eau, les lignes à haute tension, tout élément d'un réseau, etc. ;
- le polygone (ou encore la tache, ou la zone) : les limites administratives, les sites industriels étendus, les nappes d'eau souterraines, etc.

FIGURE 1

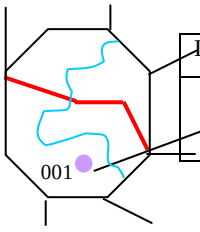
Les objets géographiques



Les points, les lignes et les polygones constituent les "couches" d'information géographique qui peuvent être superposées très facilement dans un logiciel de SIG. Pour toute donnée géographique, il est nécessaire de connaître un minimum d'information sur la donnée elle-même. Ces informations sont consignées dans les "métadonnées", elles renseignent notamment l'utilisateur sur la date de constitution, la projection géographique, etc. Les données géographiques sont définies par deux composantes indissociables : la dimension graphique et la dimension attributaire de l'information (les caractéristiques de l'objet). Les objets géographiques définis plus haut constituent la dimension graphique ou encore spatiale de l'information, et la dimension attributaire des données est renseignée dans des tables indissociables des objets graphiques dans le SIG. Par un mécanisme de jointure et grâce à un identifiant unique qui sert de clef, des informations sanitaires, environnementales, démographiques peuvent être jointes aux informations intégrées dans le SIG.

Objets Géographiques : données spatiales organisées en "couches"

Données attributaires : données alphanumériques structurées en bases de données



Identifiant	Activité	Date de début d'activité	Date de fin d'activité	Production annuelle (t)
001	Industrie plastique	1959	1999	xxx

Par exemple, une industrie (point) est associée à un identifiant qui permet de joindre l'objet géographique ponctuel aux informations attributaires qui le caractérisent, comme une date de début et de fin d'activité, une production annuelle, etc. De la même façon, lorsque l'on dispose de données de santé géoréférencées à une maille administrative donnée – prenons l'exemple d'un taux de mortalité par localisation cancéreuse par commune – il est très simple d'associer ces données à la couche géographique des communes (polygones). Les limites communales constituent le fond de carte du SIG. Les données de santé pourront être cartographiées et croisées spatialement avec d'autres données visant à caractériser l'exposition environnementales. Les données peuvent être intégrées et représentées dans un SIG a partir du moment où elles disposent d'une référence spatiale comme des coordonnées géographiques (x,y) ou un identifiant commun avec des objets géographiques pour lesquels il existe déjà une couche graphique (pour la couche des communes, le code dit "code INSEE").

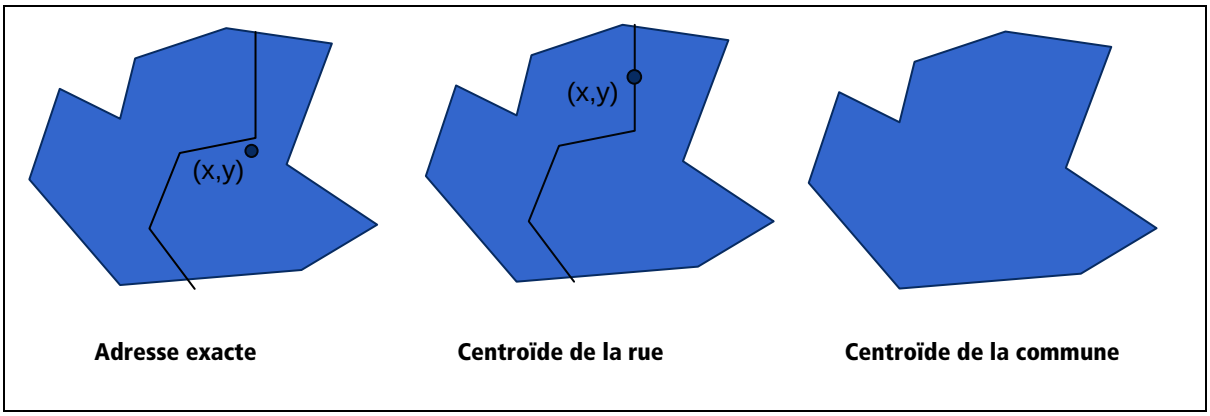
L'opération qui consiste à associer à une adresse postale un couple de coordonnées (x,y) s'appelle le géocodage. En fonction des objectifs de l'étude et les données disponibles, la précision du géocodage est variable (figure 2) : géocodage à l'adresse exacte, à la parcelle cadastrale, à la rue, à la commune, etc. Lorsque l'on fait un géocodage à la commune, cela signifie que les coordonnées de l'objet géoréférencé sont ceux du centroïde de la commune.

Un objet géographique possède donc une double caractéristique : une référence spatiale pour sa localisation et des données attributaires.

La configuration des données d'une étude conditionne ainsi la mise en place ou non d'un travail SIG. Lorsqu'une étude épidémiologique se met en place et que sa composante spatiale nécessite la création d'un SIG, il est, de ce fait, impératif de s'interroger sur les données disponibles et à mobiliser.

I FIGURE 2 I

Les géocodages



2.2.2 Que faut-il savoir à propos des données ?

Plusieurs questions doivent être élucidées préalablement à la mise en œuvre d'un SIG. Ces questions ne sont pas différentes de celles que l'on doit nécessairement se poser en préalable à toute étude. Néanmoins, la dimension géographique de l'analyse implique quelques questionnements spécifiques : de quels types de données disposons-nous ? Les données de santé et les données environnementales au sens large (occupation du sol, données sociodémographiques, infrastructures industrielles, etc.) sont-elles géoréférencées ? Si elles ne le sont pas, quels sont les moyens à mettre en œuvre pour pouvoir les intégrer au SIG ? Sont-elles disponibles ? Gratuites ou non ? À quelle échelle ? Quelle est la résolution et la qualité de ces données ? Sur quelle période sont-elles disponibles ? Quelles données va-t-on créer pour l'analyse et à partir de quelles sources de données existantes et en mobilisant quelles connaissances géographiques du contexte de l'étude (indicateurs d'exposition, tiers facteurs, etc.) ? Quelle sera la qualité des proxys ainsi obtenus ?

2.2.3 Comment choisir l'unité spatiale de référence d'une étude ?

Au vu de la diversité, de la disponibilité et de la qualité des données, il est ensuite indispensable de s'interroger sur le choix de l'unité spatiale de référence pour l'étude. L'unité spatiale choisie pour l'étude doit avant tout être pertinente vis-à-vis de la problématique traitée et des objectifs de l'étude. Le choix de l'unité spatiale de l'étude va également dépendre de la résolution spatiale des données et de leur compatibilité.

En passant d'une perspective individuelle (échelle locale ou grande échelle) à une perspective populationnelle (échelle départementale, régionale, nationale ou petite échelle), on augmente l'hétérogénéité à l'intérieur de l'unité spatiale et l'on peut plus aisément caractériser des groupes d'un point de vue socioculturel par exemple (il est plus aisé de caractériser un quartier ou une région qu'un individu par rapport à un autre). En parvenant, par le choix de la maille spatiale de référence, à faire ressortir des différences d'habitudes alimentaires, on peut par exemple parvenir à expliquer des différences observées de l'état de santé des populations (problématique des maladies cardio-vasculaires, par exemple). Dans d'autres situations, c'est la finesse des données d'exposition qui conditionnera le choix de la maille de référence pour une étude. En effet, si l'on dispose par exemple de modélisations et de mesures des rejets d'un ou plusieurs sites industriels dans le but de caractériser l'exposition à ces industries, il sera nécessaire de choisir une maille de référence qui permette au mieux de maintenir une importante variabilité entre les unités spatiales, tant pour ce facteur d'exposition principal que pour les autres indicateurs à intégrer dans l'analyse, en particulier les données sanitaires.

L'échelle de représentation et d'analyse doit donc être choisie avec précaution, pour décrire un fait de santé ou des facteurs de risque et a une influence directe sur le portrait sanitaire que l'on fait d'une population, tant au plan géographique que statistique [21].

D'un autre côté, les études écologiques géographiques trouvent leur principale limite dans le fait qu'il n'est pas possible de prendre en compte les spécificités locales dans l'analyse. Les études à l'échelle d'un groupe d'individus agrégés spatialement émettent l'hypothèse que l'exposition est répartie de manière homogène sur toute la zone. Cela peut poser problème dans la mesure où l'on ne peut pas faire référence au niveau individuel d'exposition.

2.2.4 Qu'attendons-nous des SIG ?

Les objectifs de l'étude vont déterminer à quelles étapes de l'étude il sera nécessaire de créer un SIG et à quelles fins. Les spécialistes des sciences de l'information géographique maîtrisent les possibilités techniques du SIG et c'est en faisant le lien entre leurs connaissances, à la fois techniques et contextuelles – connaissances des territoires notamment – que les objectifs de la mise en place du SIG dans une étude pourront être définis.

Lorsque l'on n'est pas *a priori* familier des SIG et des méthodes d'analyse spatiale, les attentes que l'on peut avoir vis-à-vis des SIG peuvent, dans certains cas, se restreindre à la cartographie simple d'une situation ou, tout au contraire, dépasser le seul champ d'action des SIG.

Le présent document s'attache justement à clarifier les possibilités qu'offre un SIG dans une étude épidémiologique à forte composante géographique.

Un SIG est un outil double :

- **un outil d'analyse** qui permet d'effectuer des traitements géographiques (création de zones tampon, intersections, croisement de couches, etc.). Le SIG intervient au même titre qu'un outil statistique dans une étude et doit par conséquent être prévu dès le départ dans son design. Son utilisation comme outil d'analyse est illustré par des exemples concrets détaillés dans le chapitre 2.4 ;
- **un outil de communication au sens large** pour la cartographie descriptive :
  - visualisation rapide des données tout au long de l'étude**.** Lorsque l'équipe projet cherche à comprendre le contexte général d'une étude, il est souvent pertinent de regarder une carte contextuelle représentant les éléments de lecture d'un territoire comme l'occupation du sol, la topographie, l'hydrographie, les densités de population, l'implantation industrielle, etc.,
  - classification et hiérarchisation des données,
  - cartographie des informations pour communiquer des résultats.

Les règles de représentation cartographique essentielles pour une communication efficace sont décrites ci-après.

2.3 La cartographie : quelques règles de représentation des données géographiques

"La carte est une représentation conventionnelle, plane, en positions relatives, de faits concrets ou abstraits localisables dans l'espace." (Comité français de cartographie).

La carte est une représentation visuelle qui donne à voir une ou des informations localisées dans un espace ainsi que les interactions éventuelles et les relations entre ces phénomènes. Elle utilise un ensemble de modalités qui relèvent d'un langage spécifique, un langage graphique, fondé sur la perception visuelle. Il en découle que la réalisation d'une carte nécessite une réflexion sur l'information que l'on veut transmettre, la nature du message qu'impliquent les choix de représentation et l'interprétation qui pourrait en être faite par le lecteur, la nature du public à qui elle est destinée, etc. Tout cela implique donc la connaissance d'un certain nombre de règles à suivre dans la réalisation d'une carte, règles qui relèvent de la sémiologie graphique – le sens que notre œil associe aux objets graphiques. Mais cette connaissance des règles ne se suffit pas à elle-même, et une bonne cartographie associe presque toujours à l'application de ces règles un souci de l'esthétique et de l'équilibre visuel qui rentrent aussi en ligne de compte dans la mesure où l'œil du lecteur bien souvent ne sera pas attiré par un document cartographique disharmonieux.

2.3.1 Éléments de sémiologie graphique

• Qu'est-ce que la sémiologie graphique ?

Notre œil, intuitivement, associe un sens aux objets graphiques qu'il perçoit. La sémiologie graphique, par l'étude de ces objets graphiques, du sens qu'ils portent, et de l'évolution dans l'histoire, a permis la définition d'un "ensemble de règles qui permettent l'utilisation d'un système graphique de signes pour la transmission d'une information" [23]. Par le respect de ces règles, le cartographe s'assure en grande partie d'une bonne compréhension de l'information qu'il représente et du message qu'il veut transmettre.

• Qu'est-ce que le langage cartographique ?

"C'est une forme d'expression dont les signes graphiques élémentaires (le point, le trait, la tache) seraient l'alphabet, dont le vocabulaire est fait de variables visuelles et dont la syntaxe est définie par les règles de la perception visuelle. (…) Le langage cartographique regroupe ainsi l'ensemble des moyens graphiques qui permettent de différencier, de comparer, d'ordonner, de mémoriser les informations transcrites sur le plan ou la carte." [23].

La cartographie est donc une discipline complexe où le cartographe associe des signes pour former des figurés en fonction de variables visuelles (figure 6). Ces variables visuelles sont les suivantes, déclinées sur les trois objets graphiques disponibles pour le cartographe, à savoir le point, la ligne ou le trait, la tache ou la zone, qui figurent en fait ce que l'on appelle en cartographie **l'implantation** (ponctuelle, linéaire ou zonale) d'un objet.

Les variables visuelles

› *La forme/la texture*

Géométrique ou symbolique, la forme est uniquement différenciatrice, c'est-à-dire qu'elle ne permet de transcrire qu'une information qualitative. Les variations de la forme ne peuvent être utilisées pour traduire un ordre ou des quantités. La forme sert à différencier des informations en implantation ponctuelle ; elle peut également être utilisée en implantation linéaire, elle permet alors par exemple de figurer des réseaux de nature différente. En implantation zonale, on l'utilise en faisant varier un figuré à l'intérieur de la zone, on parle alors de la variable visuelle texture.

Dans les faits aujourd'hui, la texture est moins utilisée car elle charge visuellement la carte beaucoup plus que la couleur, qui a les mêmes caractéristiques et que les modalités de publication et d'impression actuelles rendent plus **accessibles. Pour représenter les modalités d'une information qualitative, en implantation ponctuelle, on privilégiera la variation de forme ; en implantation zonale, on privilégiera la variation de couleur.**

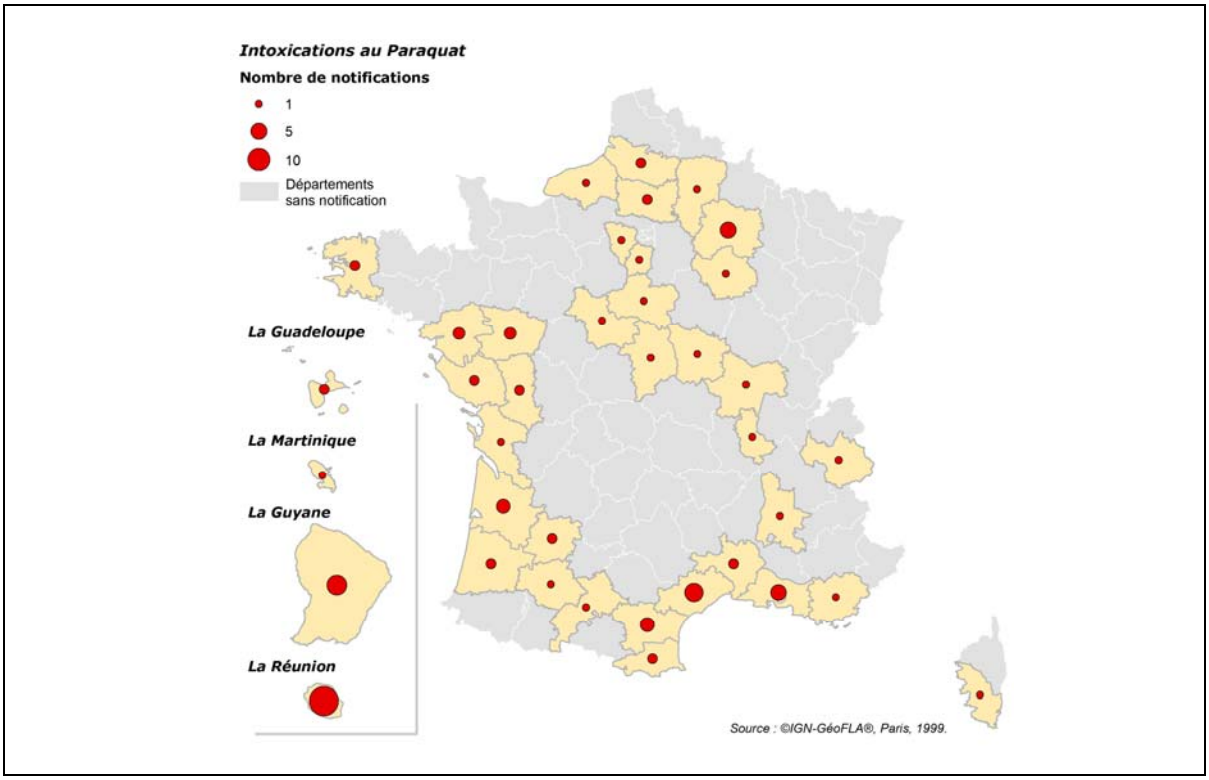
Conseil : les symboles figuratifs sont à proscrire si la densité de points à représenter est élevée ou à une échelle trop petite (France entière) car la lisibilité sera médiocre. En revanche, les symboles figuratifs peuvent convenir à une représentation à grande échelle (échelle de la commune par exemple), à condition toujours que leur densité ne soit pas trop élevée et que les symboles ne soient pas déclinés en de trop nombreuses modalités.

› *La taille*

La taille d'un objet est définie par sa longueur ou sa hauteur, sa surface ou son volume. Les variations de taille sont facilement perçues sur une carte et identifiées à des différences quantitatives. C'est la seule variable visuelle qui traduit directement une variation de quantité. Elle ne s'applique qu'en implantation ponctuelle (on fait théoriquement varier la surface du symbole proportionnellement à la quantité représentée), ou en implantation linéaire (on fait varier l'épaisseur du trait). On peut néanmoins appliquer une variation de la taille d'un symbole ponctuel pour représenter la variation d'une quantité (par exemple, la population d'une commune) en figurant le symbole proportionnel sur le barycentre (dans le SIG, le "centroïde") du polygone figurant les contours de la commune.

FIGURE 3

Variation de la taille des symboles



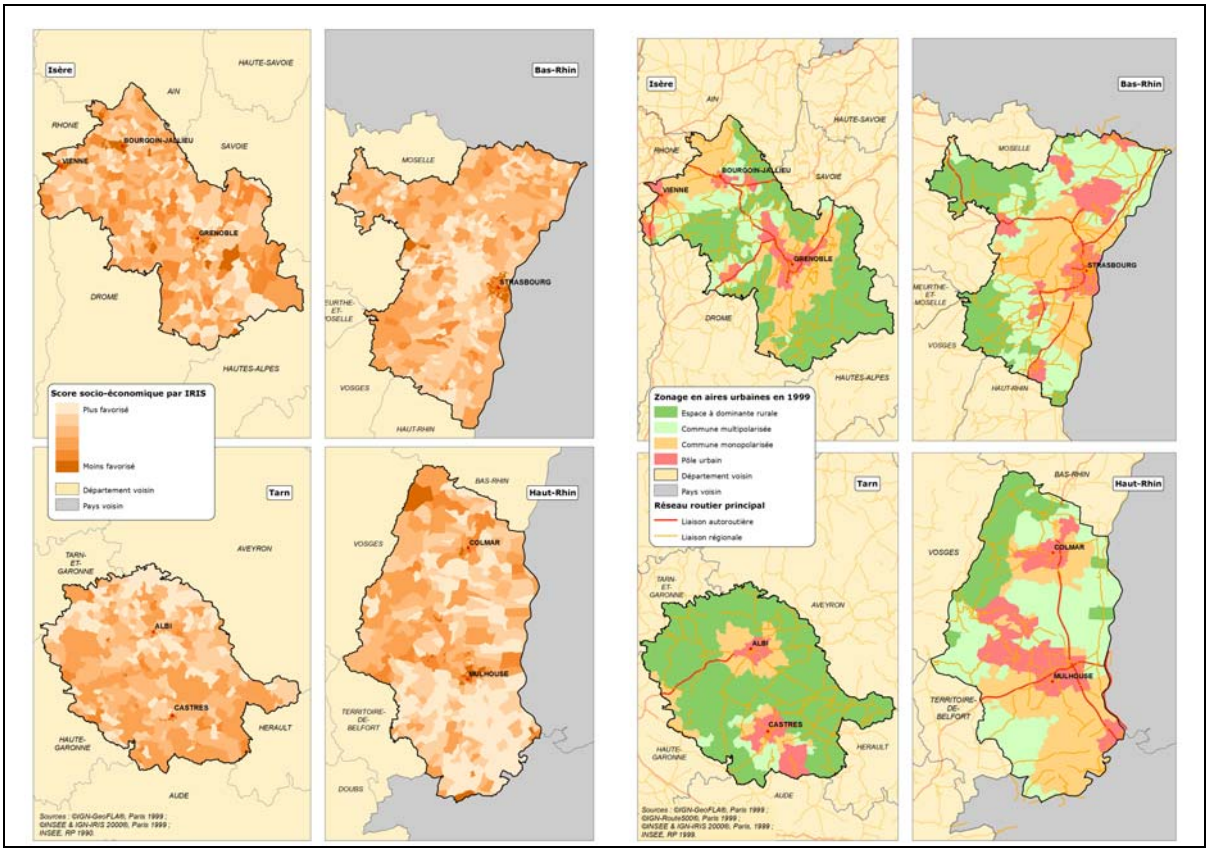
› *La couleur*

Pour Jacques Bertin, géographe et cartographe, la couleur "exerce une indéniable attraction psychologique. …[Elle] retient l'attention, multiplie le nombre de lecteurs, assure une meilleure mémorisation et en définitive, augmente la portée du message."[24]. La couleur dispose d'un fort pouvoir différentiel et d'une grande valeur esthétique. La variation de la couleur permet de figurer les variations des modalités d'une information qualitative. C'est la variable visuelle qui permet le mieux de séparer des figurés cartographiques représentant des objets de nature différente et qui transcrit avec le plus d'efficacité l'information qualitative. Dans le domaine de la santé où la visualisation de données et la cartographie servent à communiquer au public sur des sujets parfois sensibles, le choix des couleurs est important : on fera, par exemple, une utilisation modérée du rouge, qui reste associé au danger. Une fois acquis le respect d'un certain nombre de principes, une bonne utilisation de la couleur relève plus de la pratique que de l'application des règles strictes.

Attention, la variation de couleur (du bleu, du vert, du jaune, du rouge…) ne doit pas être confondue avec une variation de valeur dans une même couleur, que l'on appelle communément un "dégradé" allant, par exemple, du beige très pâle au marron foncé. La variation de valeur dans une gamme colorée (simple gamme – on fait varier la valeur pour une seule couleur – ou double gamme dans certaines représentations) traduit une information ordonnée.

| FIGURE 4 |

Valeur et couleur

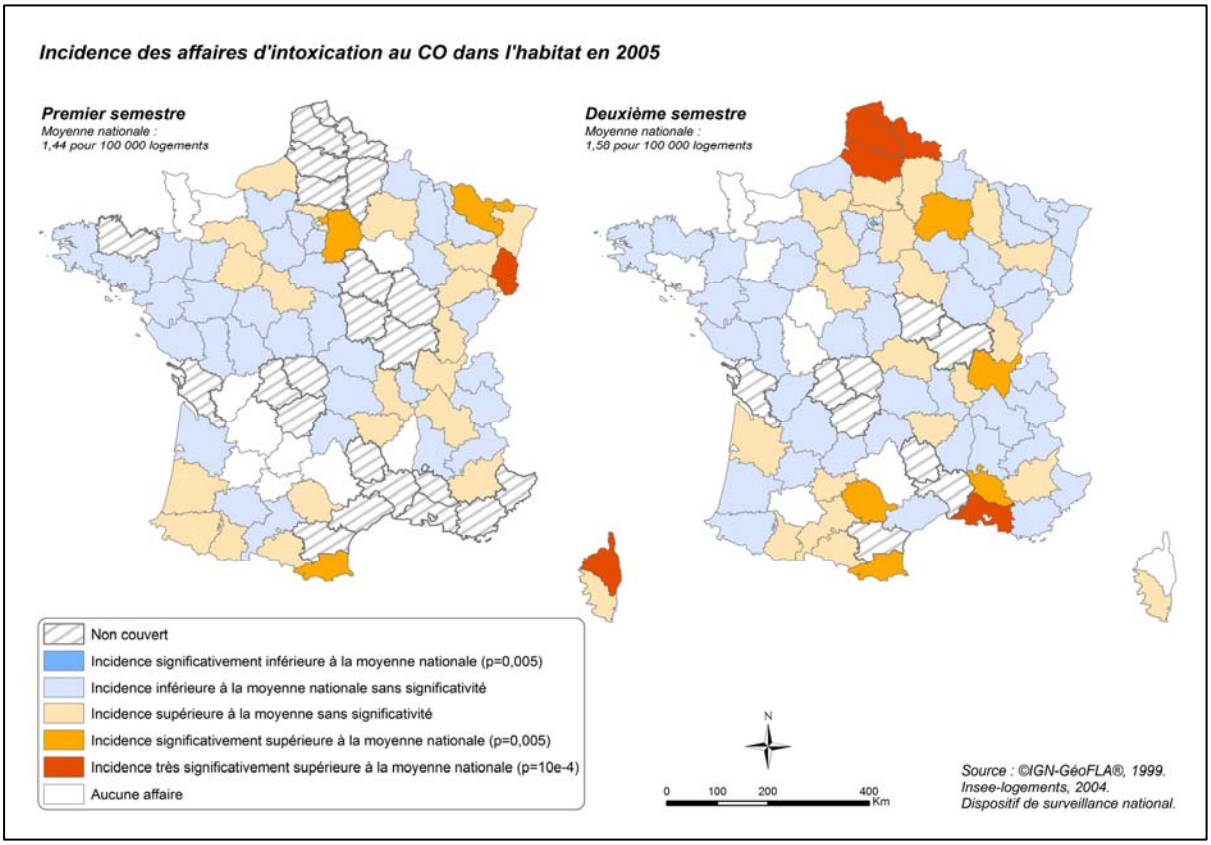


› *La valeur*

On appelle valeur le rapport entre les quantités de noir et de blanc perçues dans une surface donnée. Cela concerne aussi la couleur que l'on peut éclaircir ou foncer par apport de blanc ou de noir (voir ci-dessus). On obtient alors un "dégradé" de couleur. La valeur est une variable visuelle qui permet de traduire un ordre (utilisée uniquement pour représenter une information ordonnée), car l'œil classe naturellement les taches grisées de la plus claire à la plus foncée en associant aux taches claires les valeurs faibles et aux taches foncées les valeurs fortes. Le plus souvent, on utilise la variation de valeur sur une simple gamme, de noir et blanc ou de couleur. Cependant, il peut parfois être utile d'introduire une deuxième gamme, le saut d'une gamme à l'autre servant alors à faire apparaître une valeur seuil que l'on veut isoler ou distinguer : on peut vouloir isoler une valeur de référence, comme par exemple, en météo, sur des cartes de température, on isole le 0 °C ; on peut aussi avoir besoin de faire ressortir une valeur de la distribution statistique de la variable que l'on cartographie : moyenne, médiane, etc.

| FIGURE 5 |

Double gamme de couleurs inversées



› *Le grain*

La variation de grain s'obtient par agrandissement ou réduction d'une texture. Elle correspond à une variation de taille de l'élément constitutif de la trame. La variable grain permet de représenter des caractères ordonnés. Cependant, l'expérience montre que ce classement est limité à trois ou quatre paliers au maximum. Elle est le plus souvent délaissée au profit de la valeur.

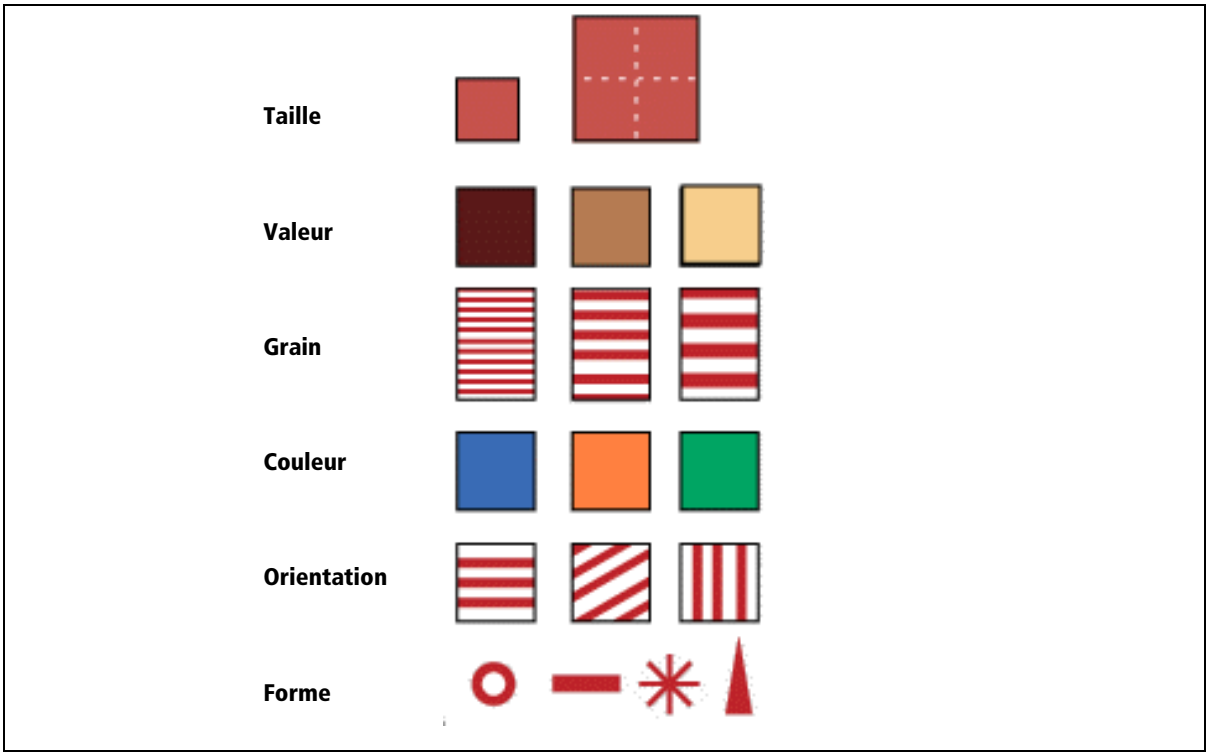
› *L'orientation*

Elle s'applique à un figuré linéaire (type hachure). Cette variable est différenciatrice et utilisée pour représenter les modalités de caractères qualitatifs.

La figure 6 récapitule les différentes variables visuelles disponibles pour la représentation cartographique.

| FIGURE 6 |

Les variables visuelles



2.3.2 Cartographier les informations : les choix de représentation

On distingue deux grandes familles de cartes : les cartes d'inventaire et les cartes thématiques. Les cartes d'inventaire, peu utilisées en santé publique, sont des cartes qui recensent un maximum d'informations sur un sujet donné, avec un objectif d'exhaustivité. Ce sont, par exemple, les cartes que l'on trouve habituellement sur les plaquettes des offices du tourisme, ou encore les cartes IGN Top25… La réalisation de ce type de cartes pose de nombreux problèmes (densité de l'information, sélection de cette information, précision, esthétique) qui lui sont assez particuliers. Elle ne sera pas abordée ici dans la mesure où elles ne font que rarement partie des cartographies nécessitées par les travaux réalisés à l'Institut, à la différence des cartes thématiques qui, elles, sont monnaie courante dans nos travaux. Les cartes thématiques sont des cartes géographiques illustrant, par l'utilisation de divers paramètres graphiques (couleur, symbolique, taille, etc.), le comportement d'un phénomène en relation avec sa localisation spatiale.

Il existe différents types de cartes dites "thématiques" : cartes par symboles proportionnels, cartes choroplèthes (représentation cartographique d'une information quantitative par plages colorées), cartes isoplèthes (zones délimitées par des courbes d'iso-concentrations, de niveaux, de températures, etc.), cartogrammes (la taille des unités spatiales varie en fonction des valeurs représentées, ce sont des schémas plus que des cartes, et ils ne sont absolument pas envisageables avec des outils SIG standards), etc.

Le choix de la méthode de cartographie à adopter va donc dépendre essentiellement de la nature de l'information que l'on souhaite représenter, et des caractéristiques de la ou des variables qui portent cette information. En fonction des variables à représenter, le choix se portera sur tel ou tel type de cartographie thématique, et sur telle ou telle variable visuelle – ou combinaison de variables si l'information à représenter sur une même carte est multiple.

Les variables à cartographier peuvent être regroupées en quatre grandes catégories :
- les variables qualitatives nominales : l'information concerne des catégories, des types ; elle est destinée à classer les entités sans notion d'ordre (type de culture, par exemple, pour une carte sur l'agriculture, occupation du sol…) ;
- les variables qualitatives ordonnées : l'information concerne là aussi des catégories ou des types mais contient une notion d'ordre : par exemple, le type de route – entre chemin vicinal, route départementale, autoroute – sous-tend une notion d'ordre : le trafic est plus important sur une autoroute que sur un chemin vicinal ; ou encore la caractérisation d'une ville selon son statut administratif : entre une commune simple, un chef-lieu de canton, une

sous-préfecture, une préfecture, une préfecture de région, ou encore une capitale nationale, il existe un classement par ordre d'importance ;
– les variables quantitatives brutes, ou effectifs : ce sont toutes les variables qui dénombrent une quantité (une population par exemple) dans l'absolu ;
– les variables quantitatives relatives, qui expriment un rapport : la quantité étudiée est rapportée à une autre quantité : densité (population/surface), taux (population particulière/population totale)…

TABLEAU 1

Représenter une information sur une carte

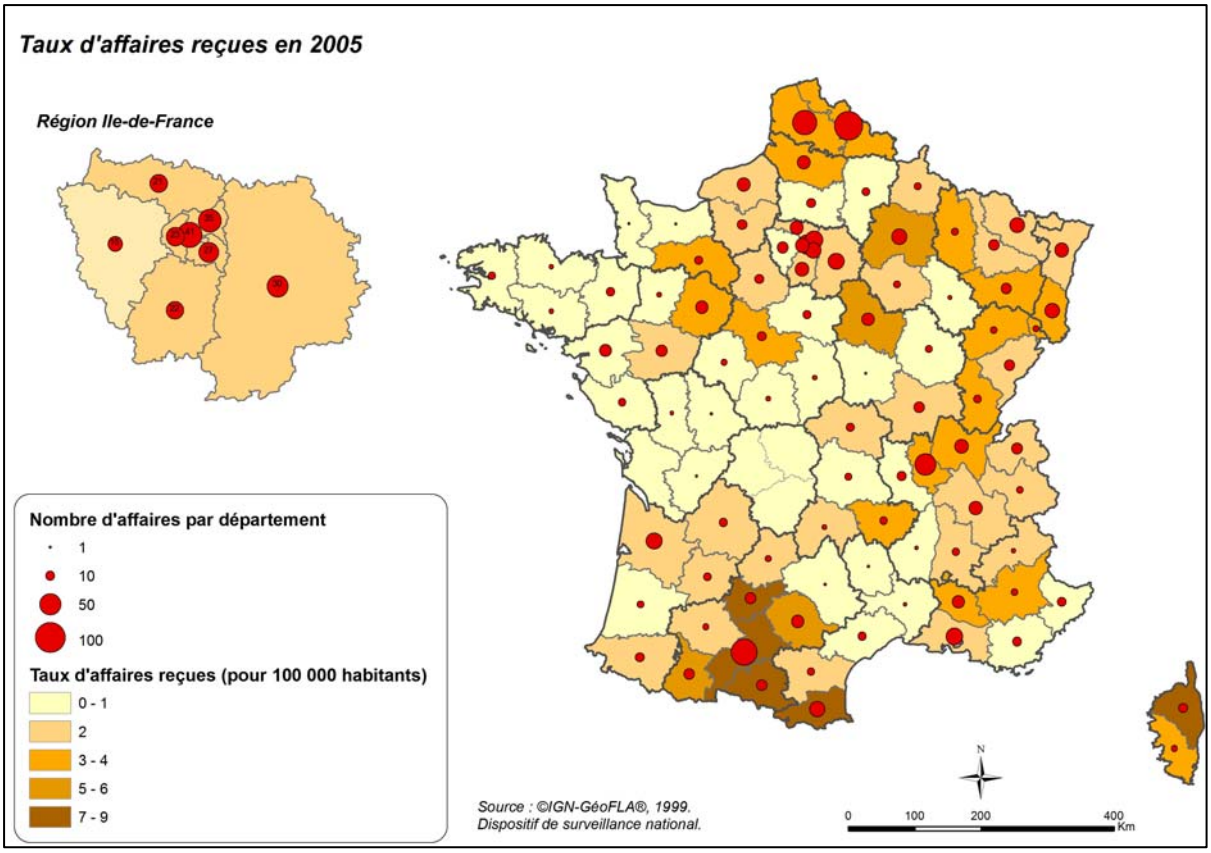
Données		Implantation géographique		
		ponctuelle	linéaire	zonale
Qualitatives	Nominales	**Forme** et/ou **couleur**	**Forme** et/ou **couleur**	**Couleur** et/ou texture-structure (variation de la forme sur une trame remplissant la surface)
	Ordonnées	**Taille**	**Taille**	**Valeur**
Quantitatives	Effectifs (population…)	**Taille** avec une variation proportionnelle à l'effectif représenté	**Taille**	**Taille** sur un objet en implantation ponctuelle placé au barycentre (ou encore centroïde) de la surface concernée. JAMAIS d'utilisation de la valeur pour un effectif.
	Rapport (taux…)	**Valeur** si possible (parfois la taille mais elle est à préférer pour représenter des effectifs)	**Valeur** si possible (parfois la taille mais elle est à préférer pour représenter des effectifs)	**Valeur**

Le tableau précédent (tableau 1) est une bonne synthèse des règles à appliquer dans le choix de la variable/des variables visuelle(s) à utiliser en fonction de l'information que l'on souhaite cartographier.

Le tableau se lit de la manière suivante :
– en ligne : le type de données que l'on souhaite cartographier ;
– en colonne : sur quels objets géographiques (on parle d'implantation) porte l'information à représenter ;
– dans les cases figure la ou les variables visuelles à utiliser. On n'utilise jamais la variable visuelle valeur (dégradé du blanc au noir, ou dégradé dans une même gamme colorée) pour représenter une variable quantitative brute (effectif, population…) à cause de l'effet de taille qu'elle entraîne : ainsi, sur des unités géographiques de tailles très hétérogènes (les départements d'Ile-de-France, par exemple, ainsi Paris est beaucoup moins étendu que la Seine-et-Marne), les grandes surfaces ressortiront beaucoup plus visuellement que les petites, alors que la population des plus petites pourra être bien plus importante (le cas de Paris et du reste de l'Ile-de-France, pour ce qui est de la population, est un très bon exemple). Par contre, cet effet de taille disparait si l'on introduit la notion de rapport : ainsi, par exemple, pour la cartographie de la densité de la population, puisque la surface est prise en compte dans le calcul de la variable même, l'effet de taille disparaît et il devient tout à fait adapté d'introduire la valeur comme variable visuelle pour ce type d'information ;
– **représentation cartographique de plusieurs variables : carte combinant plusieurs variables ou collection de cartes ?** Il est parfaitement possible de cartographier plus d'une information sur une même carte thématique, à condition toutefois que le regroupement sur une même carte ait un sens (figure 7). Il faut néanmoins tenir compte, dans ce cas, de la pertinence du regroupement sur une même carte de ces informations ainsi que de la lisibilité.

| FIGURE 7 |

Cartographier deux variables sur la même carte

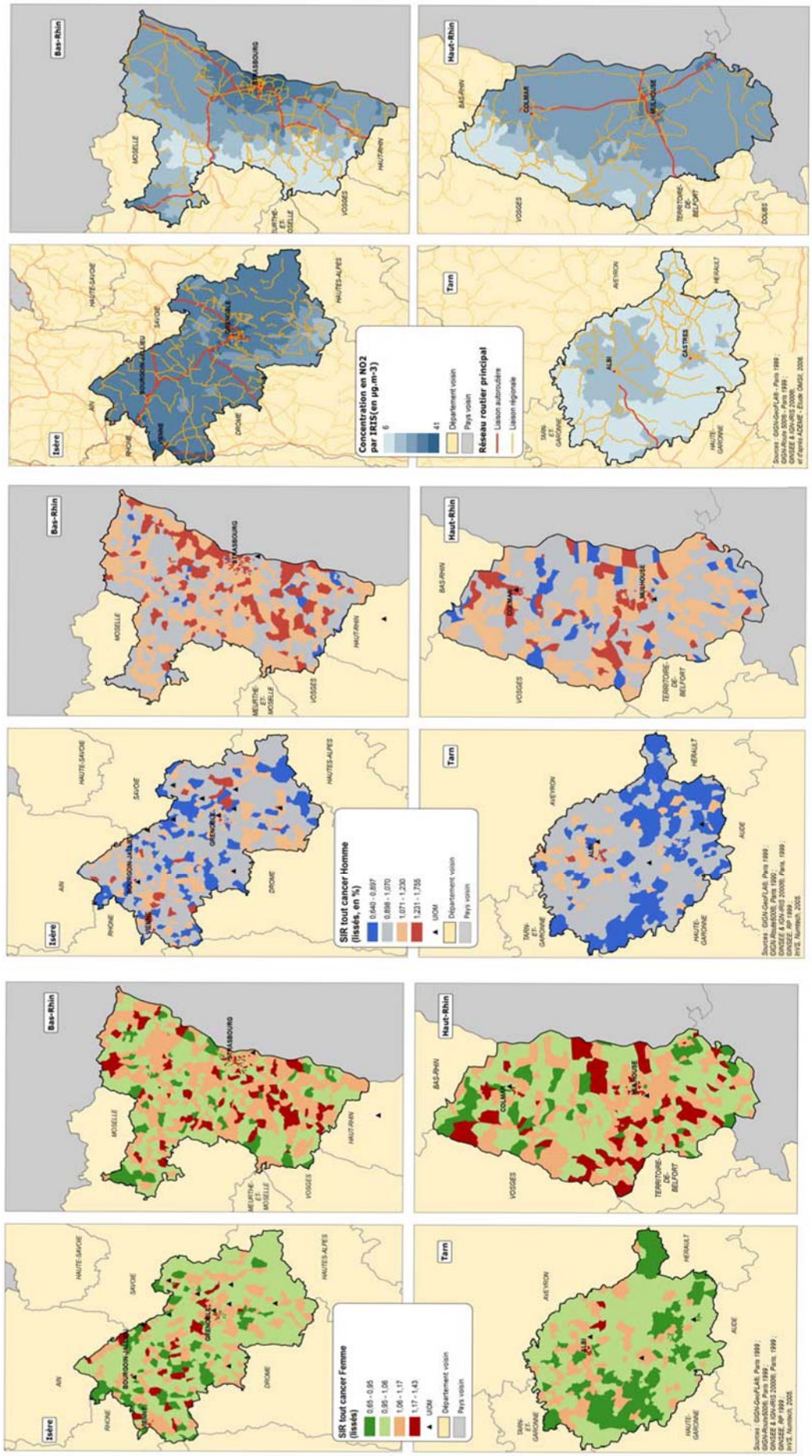


Il faut ainsi savoir qu'au-delà de deux (trois au grand maximum) variables cartographiées sur un même espace la lecture devient ardue. Il peut être alors préférable de réaliser une collection de cartes, en adoptant des modes de représentation analogues d'une carte à l'autre (figure 8).

Selon le type d'information cartographiée et les objectifs du travail, la collection de cartes pourra se décliner de différentes manières : soit une série de cartes de la même zone géographique faisant figurer les différentes variables d'intérêt, soit une série de cartes réunissant les différentes variables d'intérêt mais déclinées par régions/départements ou autre zonage géographique pertinent.

I FIGURE 8 I

Exemple d'une collection de cartes



– **la discrétisation d'une variable continue**

En dehors de la cartographie d'une variable quantitative brute représentée par la variation proportionnelle de la taille d'un symbole ponctuel, il sera souvent nécessaire de recourir à une méthode de discrétisation pour cartographier une information quantitative. Toute méthode de discrétisation pour la cartographie consiste "à subdiviser le domaine de variation d'une série statistique continue en classes de valeurs" [23]. Les méthodes sont nombreuses. Le choix dépend des propriétés de la distribution et des objectifs fixés quant à l'information à communiquer comme le montre le tableau 2. Il n'y a donc pas *a priori* de bonne ou de mauvaise méthode de discrétisation, et il n'est pas possible *a priori* d'en recommander une plutôt qu'une autre, le choix devant se faire au cas par cas.

TABLEAU 2

Quelques clefs pour choisir une méthode de discrétisation

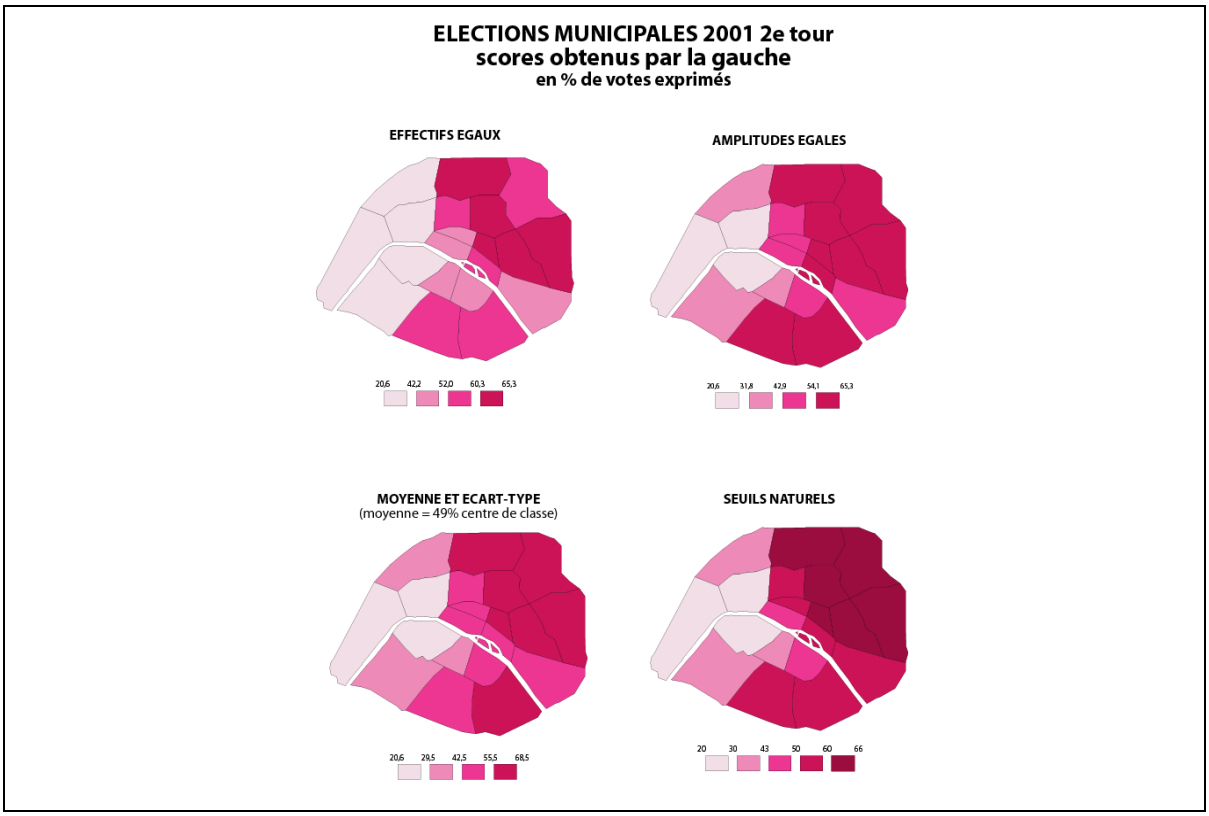
Nature de l'information		Méthode appropriée	Méthode inappropriée
Forme de la distribution	Normale	Moyenne écart-type ou classes équiprobables	Classes d'amplitude égale ou progression géométrique (à droite) ou moyenne écart-type ou classes équiprobables
	Dissymétrique	Progression géométrique (à gauche) ou méthode de Jenks	Amplitude égale/moyenne-écart-type
	Uniforme	Classes d'amplitude égale	Progression géométrique
	Plurimodale	Seuils naturels – Jenks	Amplitudes égales
Objectif recherché	Faire ressortir les valeurs extrêmes	Seuils naturels – Jenks ou comparaison à une valeur standard/de référence	Amplitudes égales
	Comparaison	Moyenne écart-type ou classes équiprobables (si distribution normale) ou discrétisation par quantiles	Arbitraire/seuils naturels/amplitude égale/Jenks
	Mettre en évidence une configuration spatiale	Jenks	Arbitraire

ArcGIS® propose un certain nombre de méthodes de discrétisation "prêtes à l'emploi", celle qu'il utilise par défaut est la méthode de Jenks, dite aussi "des seuils naturels" [25]. Mais il est également possible de fixer soi-même les bornes de ses classes, et ce d'autant plus aisément que le logiciel propose de visualiser la distribution statistique des modalités de la variable sur un diagramme de distribution. Le choix de la méthode est à adapter à la variable représentée (il peut ou non être utile d'isoler une valeur de référence, de faire ressortir ou non certains indicateurs de dispersion comme la médiane, l'écart-type…, il peut être pertinent ou non d'adopter une discrétisation par quantiles…). Cependant, il faut noter de manière générale que l'œil ne perçoit aisément les variations de valeur au sein d'une trame, qu'elle soit en noir et blanc ou en couleur, que sur sept paliers au grand maximum, et que le plus souvent, pour une bonne lisibilité, il sera préférable de se limiter à quatre ou cinq classes.

Les classes sont ensuite identifiées dans la légende par les valeurs des bornes mentionnées à côté des caissons ou des symboles qui leur sont affectés sur la carte. Le choix de la méthode de discrétisation conditionne grandement le résultat cartographique comme le montre la figure 9, la même variable étant représentée selon quatre discrétisations différentes. Ce choix n'est donc pas à faire à la légère. Pour un lecteur averti, il peut être utile de faire apparaître sur la carte le mode de discrétisation adopté.

| FIGURE 9 |

Cartes montrant les différences entre méthodes de discrétisation



Pour plus d'informations, il convient de se référer à l'article disponible à partir de l'URL : http://www.hypergeo.eu/spip.php?article374 sur les discrétisations en elles-mêmes et au document suivant, disponible à partir de l'URL : http://www.hypergeo.eu/article.php3?id_article=274 sur les cartes choroplèthes.

2.3.3 Des éléments clefs pour la lecture d'une carte

La réalisation d'une carte implique de faire figurer de manière systématique un certain nombre d'éléments indispensables, qui sont de véritables clefs de lecture de la carte elle-même. Par ailleurs, elle peut être enrichie de certains éléments dont l'ajout peut rester optionnel, on distinguera donc ces deux catégories d'éléments.

- Les éléments incontournables

› *Le titre*

Ces éléments sont tout d'abord un titre, répondant synthétiquement aux questions : quoi, où, et éventuellement, quand, concernant les informations représentées. Ces informations peuvent aussi se répartir efficacement, et pour plus de concision, entre le titre de la carte lui-même et le titre du bloc de légende.

› *La légende*

La carte doit également comprendre une légende, qui met en regard des variables visuelles utilisées et les informations représentées décrites là encore de façon à la fois complète et synthétique. Si l'information cartographiée est une variable quantitative, il est indispensable de mentionner en légende l'unité dans laquelle cette variable s'exprime ; si c'est une variable continue que l'on a discrétisée, il est nécessaire de figurer précisément les bornes des classes (précision de l'exclusion ou de l'inclusion des bornes dans les classes mentionnées…).

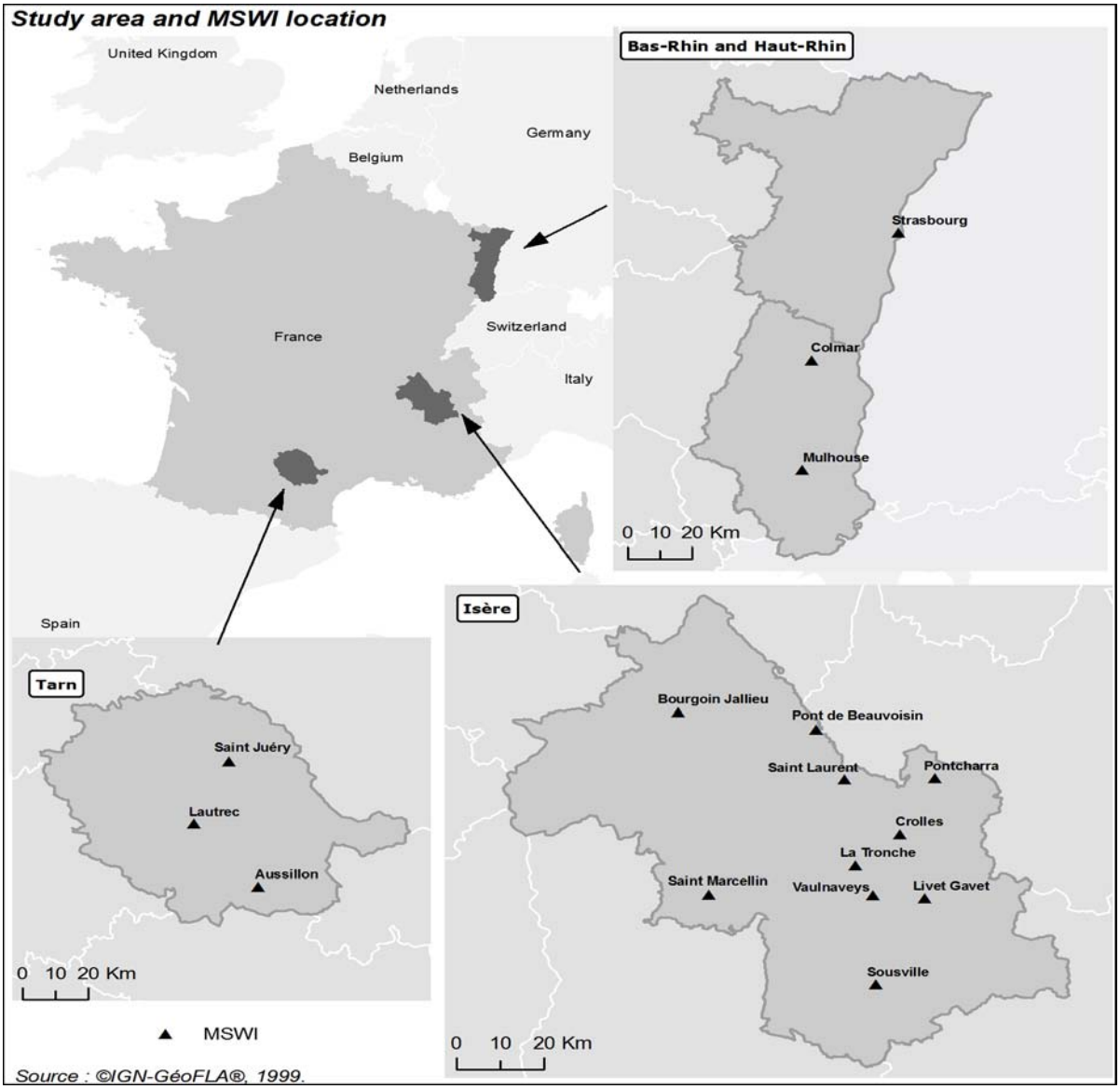
› *L'échelle*

Une carte doit également toujours comporter une échelle : une carte est en effet une représentation en réduction de l'espace que l'œil doit pouvoir en quelque sorte reconstituer grâce à la mention de cette échelle. L'échelle doit être une échelle graphique car la lecture en est bien plus immédiate que sur une échelle numérique, et elle a en outre l'avantage d'être dynamique (elle évolue automatiquement dans le SIG en fonction de la taille de la carte ou du niveau de "zoom"

choisi par l'utilisateur. Les outils SIG proposent une grande diversité d'échelles graphiques à la disposition des cartographes. Il est recommandé d'utiliser un style d'échelle le plus simple et le plus lisible possible, dont l'unité de distance est claire et dont les mentions numériques seront des chiffres ronds et non des distances peu usuelles ; il faut également faire attention à l'unité dans laquelle ces distances s'expriment (éviter le mètre pour une carte régionale ; éviter des repères numériques de 7,3 km et privilégier les valeurs rondes telles 10, 20, 50, 100…). Exceptionnellement, on peut se passer de faire figurer l'échelle : les cartes de la France métropolitaine, ou dans certains cas, des cartes par régions peuvent à la grande rigueur se passer d'une échelle car notre œil est assez habitué à sa représentation et à ses dimensions spatiales (les cartes de France du bulletin météo télévisé ne comportent pas, par exemple, de mention d'échelle car le téléspectateur est complètement familier de ces représentations et sait reconnaître sa ville sur les différents points où sont figurés températures ou symboles météorologiques). Cela dit, cela est vrai si la carte s'adresse à un public habitué à ce territoire et à ses délimitations ; dans le contexte d'un lectorat plus large, on préférera faire figurer une échelle même sur un territoire aux dimensions bien connues. Par contre, on évitera de surcharger la représentation d'une carte qui comporte à la fois la représentation d'un espace délimité (zone d'étude, département, région…) et un carton de localisation (petit encart figurant, soit la France entière, soit de façon différentielle la zone cartographiée plus en détail) : celui-ci n'est là qu'à titre très informatif et n'est destiné qu'à localiser plus globalement la zone délimitée que l'on étudie, il n'y a pas lieu d'y ajouter des repères d'échelle.

FIGURE 10

Exemple d'une carte métropolitaine sans mention de l'échelle



Enfin, quand on souhaite représenter à la fois le territoire métropolitain et l'outre-mer, il faut veiller au fait que l'échelle varie souvent beaucoup d'un territoire représenté à l'autre et garder à l'esprit que la Guyane est bien plus étendue que la Réunion, et que cela doit être mis en évidence par la présence d'une échelle.

› *La flèche du Nord*

Il est également souvent utile de faire figurer sur une carte la flèche indiquant le Nord. En effet, il arrive que l'on soit amené (assez rarement) à tourner une carte pour des raisons de lisibilité et à ne plus se conformer à la convention qui veut que le Nord géographique pointe vers le "haut" de la page. La mention du Nord devient alors indispensable. Elle reste utile, même sans rotation de la carte, quand on représente une zone plus limitée du territoire, elle est alors, au même titre que l'échelle, un élément utile à l'œil pour se repérer. Elle devient plus facultative pour la représentation à plus petite échelle du territoire métropolitain dans sa globalité, car tout-un-chacun dans cette dernière situation sait intuitivement sans avoir besoin qu'on le lui rappelle que le Nord géographique se trouve en "haut" de la page.

› *La source*

Enfin, la carte doit mentionner la source des informations qu'elle représente, de façon à la fois synthétique et exhaustive. Discrète sur la carte, elle renseigne néanmoins le lecteur sur la nature des données utilisées, leurs auteurs, leur date, autant d'éléments complémentaires nécessaires à la bonne compréhension de la carte. Il faut y faire apparaître l'organisme producteur et la date des données représentées (les variables que l'on a cartographiées) aussi bien que les informations concernant l'origine et la date des fonds de cartes.

- **Les éléments optionnels**

› *L'encart ou le carton de localisation*

Quand la zone d'intérêt n'est pas le territoire métropolitain dans son ensemble, ou un autre territoire bien connu du public auquel la carte est destinée, il peut ne pas être inutile d'ajouter à la carte un "carton de localisation", en encart de la carte elle-même : le territoire d'intérêt est alors mis en valeur sur une représentation en taille très réduite d'un territoire de référence bien connu du lectorat de la carte. Dans d'autres situations, il peut être nécessaire d'agrandir, en encart de la carte, une zone particulière du territoire : par exemple, sur de nombreuses cartographies de la France métropolitaine dans son ensemble, on ajoute un encart portant sur l'Ile-de-France, territoire aux entités spatiales peu étendues et qualifiées par des valeurs fortes pour de nombreuses problématiques ce qui compromet un peu la lisibilité sur ce territoire.

› *Les quadrillages/carroyages*

Sur certaines cartes apparaît un carroyage qui, le plus souvent, est destiné à faciliter le repérage visuel des distances. Très utilisé sur les plans de ville, ou sur les cartes à grande échelle, il n'a que peu de raisons *a priori* de figurer sur les cartes produites sur des documents épidémiologiques ou de santé publique.

› *Le logo*

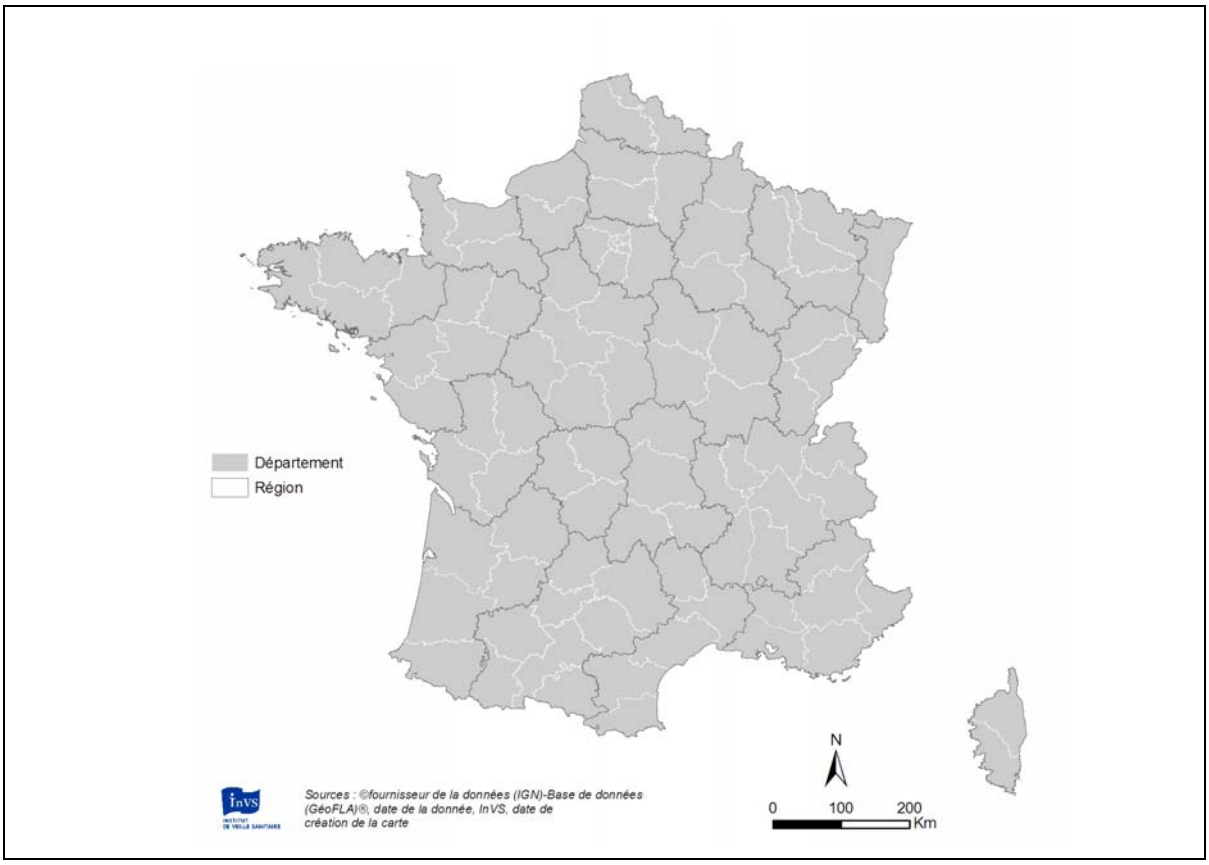
En amont de la réalisation d'une carte, il est nécessaire de se poser systématiquement la question du destinataire du document cartographique. Un document destiné au grand public ou, au contraire, à un public averti, n'implique pas forcément les mêmes choix ; on ne fera pas abstraction de cette question car elle conditionne autant le choix de la variable représentée, que la présentation du titre, le choix des variables visuelles et la présentation de la légende : vocable à utiliser, précaution à prendre par rapport à une donnée qui pourrait être sensible selon le public qui en sera destinataire, choix des couleurs pour ne pas stigmatiser une zone géographique par rapport aux autres quand la donnée cartographiée sous-entend un potentiel jugement de valeur. Il ne s'agit pas, en se posant cette question et en adaptant ses choix de cartographie à son public, de fausser l'information que l'on cartographie, mais plutôt d'être conscient que les possibilités de choix qui s'offrent au cartographe lui permettent d'ajuster au mieux les modes de représentation au message qui doit être transmis. *A contrario* d'ailleurs, il est nécessaire de garder à l'esprit que, justement, la variété des possibilités qui s'offrent au cartographe peuvent d'introduire un message trompeur, et que la déontologie du cartographe, c'est aussi de savoir toujours garder justesse et précision à l'information cartographiée.

- *L'équilibre visuel d'une carte*

Voici en exemple une carte (figure 11) au format 15x15 cm environ, soit le format dans lequel on compose généralement une carte sous ArcGIS® dans un document au format A4 puisque la représentation du territoire français s'inscrit *grosso modo* dans un carré.

| FIGURE 11 |

Exemple de mise en page d'une carte

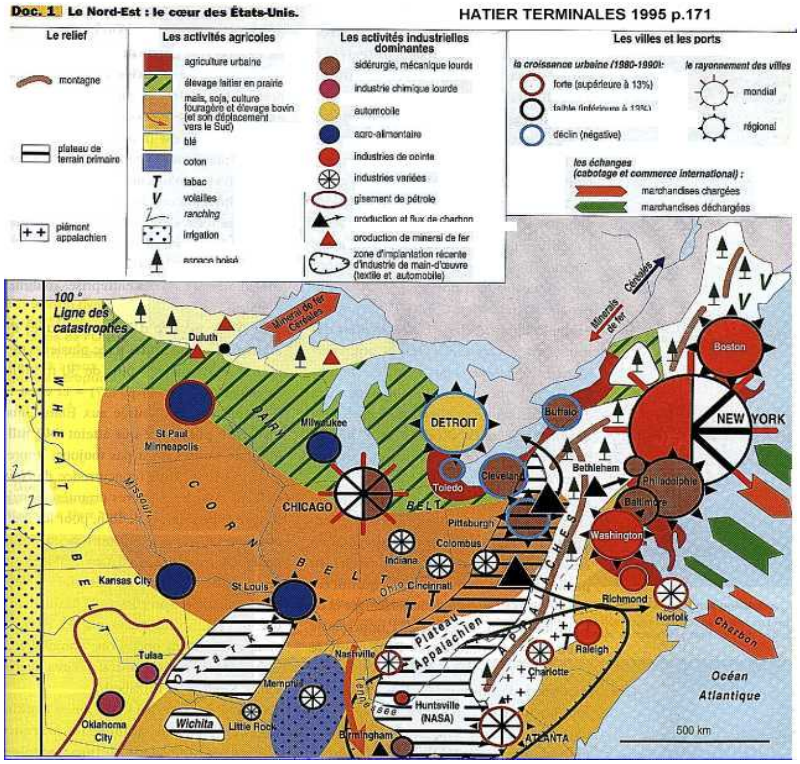


Le temps moyen de lecture d'une carte est faible puisqu'on considère généralement que l'œil ne s'attarde pas sur une carte plus de 20 secondes, ou, du moins, qu'il est indispensable que l'essentiel de l'information ait été perçue par le lecteur dans un temps inférieur à cette durée. L'expression "sauter aux yeux" est donc, ici plus qu'ailleurs, particulièrement adaptée à l'information cartographique. Tout le travail du cartographe doit donc viser cet objectif de lisibilité et de rapidité de lecture.

De ce fait, l'aspect esthétique entre nécessairement en jeu dans l'attrait que doit avoir la carte pour que l'œil s'y attarde. L'équilibre visuel y contribue grandement et dépend essentiellement de la mise en page du document graphique. Afin de créer cet équilibre visuel, il convient de construire son document comme un tout, en insérant l'ensemble des éléments clefs de la carte dans un cadre. Point n'est besoin pour cela de multiplier les cadres, justement, dans la mise en page. Les éléments constitutifs de la carte et la carte elle-même suffisent à construire cet ensemble équilibré et harmonieux. Selon la forme du territoire cartographié, on veillera ainsi à utiliser les différents "vides" pour l'insertion de la légende, de l'échelle, etc. de façon à construire ce cadre visuel. Enfin, l'œil faisant naturellement plus facilement le lien entre des éléments qui sont proches les uns des autres qu'entre des éléments éloignés, on apportera un soin particulier à intégrer les éléments clefs de lecture, en particulier la légende et l'échelle, à proximité du territoire représenté. Il faut bannir les cartes dont la légende serait à une page différente de celle de la carte, par exemple. Enfin, la lisibilité devant rester un maître mot à toutes les étapes du travail de cartographie, on privilégiera toujours l'allègement de la carte, quitte à réaliser plusieurs cartes, plutôt que d'essayer à tout prix de représenter toutes ses informations sur la même carte, en la surchargeant jusqu'à la rendre difficile, voire impossible à lire (figure 12).

| FIGURE 12 |

Exemple d'une mauvaise carte (carte extraite d'un manuel scolaire de terminale histoire-géographie, 1995)



La figure 12 présente un bon exemple de carte difficile à lire. À vocation synthétique, elle se révèle en fait surchargée et à la limite de la lisibilité. L'œil a du mal à percevoir le territoire cartographié tant la densité des figurés et variables visuelles utilisées (cercles, flèches, textures, symboles ponctuels en tout genre, couleurs…) est importante. De plus, comme on a cherché à tout représenter sur une même carte, il n'apparaît pas de hiérarchie entre les informations, le titre est extrêmement général et donc très peu accrocheur.

Enfin, pour conclure, nous reposerons, la question du public destinataire de ce document. Il conditionnera lui aussi énormément le choix des informations figurées, le choix des formulations (comment l'information est présentée), le choix des variables visuelles (les couleurs notamment)… Même en respectant scrupuleusement les règles de sémiologie graphique et de cartographie présentées par ce document, il n'y a pas une manière unique de représenter ces informations sur une carte. Il convient, en résumé, de toujours veiller à ce que le message soit correctement perçu par le destinataire, sans déformation, en veillant à la façon dont il peut être interprété, mais dans le respect de l'information elle-même, etc. Un juste dosage en quelque sorte, entre déontologie et pédagogie, qui rappelle que la cartographie va souvent bien au-delà de la seule maîtrise de règles et d'outils ou de logiciels, et que c'est une véritable discipline à part entière. Au-delà de la cartographie qui est une des disciplines majeures de la géomatique, on utilise l'éclairage géographique et les systèmes d'information géographique pour décrire un contexte d'étude et la répartition spatiale des faits de santé.

2.4 Les SIG comme outil d'analyse descriptive : étude des relations spatiales entre les entités géographiques

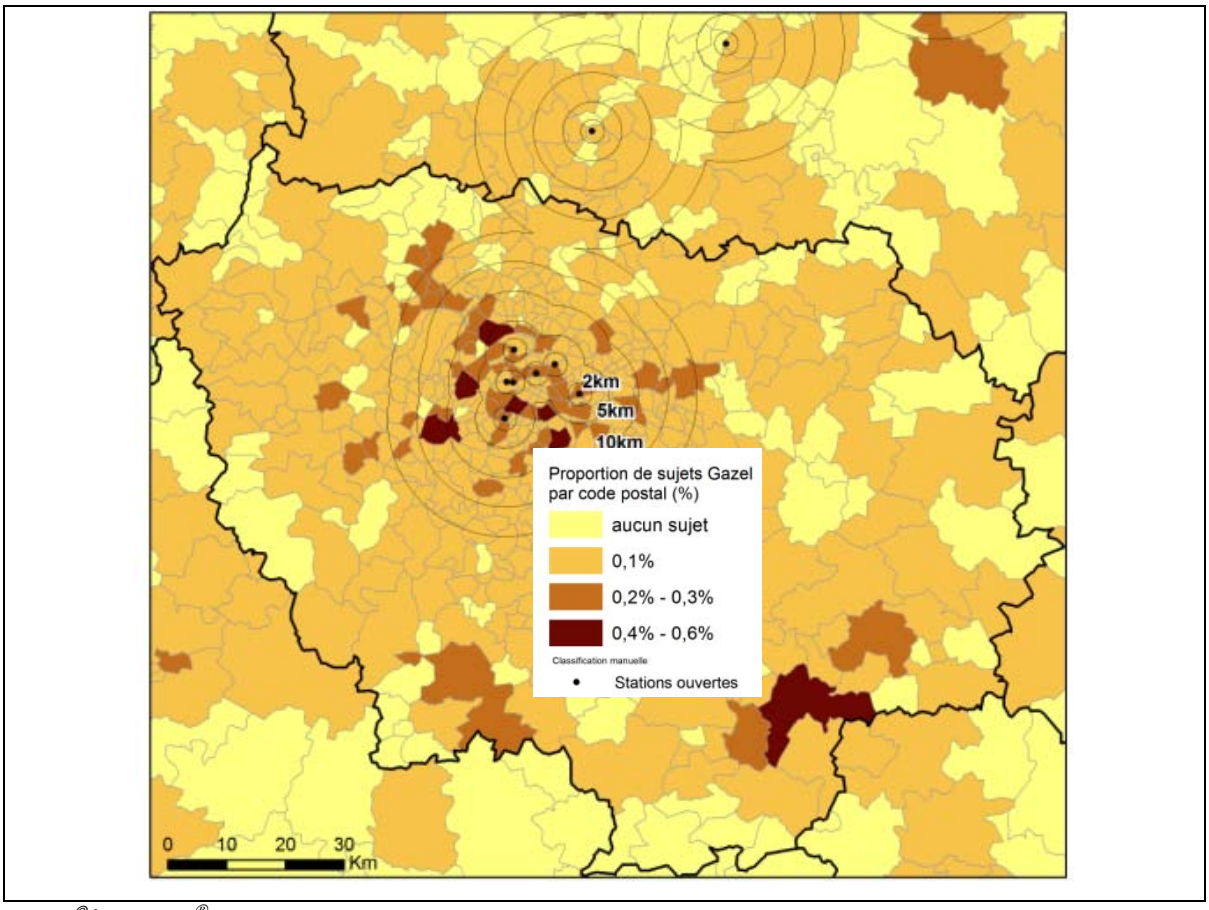
Comme il l'a été dit dans le chapitre 2, le SIG est un outil de communication mais aussi un outil d'analyse. Il offre quantité d'options pour la description et l'analyse de l'information géographique. Les traitements géographiques les plus utilisés se retrouvent aujourd'hui dans tous les logiciels de SIG. Il s'agit, par exemple,

d'opérations comme l'intersection spatiale de deux couches d'information géographique (on va découper une couche "département" par exemple au moyen d'une couche "commune") ou encore le découpage d'une couche d'information à partir d'une entité spatiale, le calcul de distance et de surfaces, etc., qui permettent finalement de créer de nouvelles informations.

Parmi les traitements géographiques de base, la construction de zones tampon autour d'entités spatiales est couramment utilisée pour mesurer la distance d'un sujet ou d'une entité administrative (commune, code postal) à un site pollué par exemple ou encore à une station de mesure de la qualité de l'air (figure 13).

I FIGURE 13 I

Zones tampons ou cercles concentriques (buffers en anglais)



Source : © IGN-GéoFLA ®, 1999.

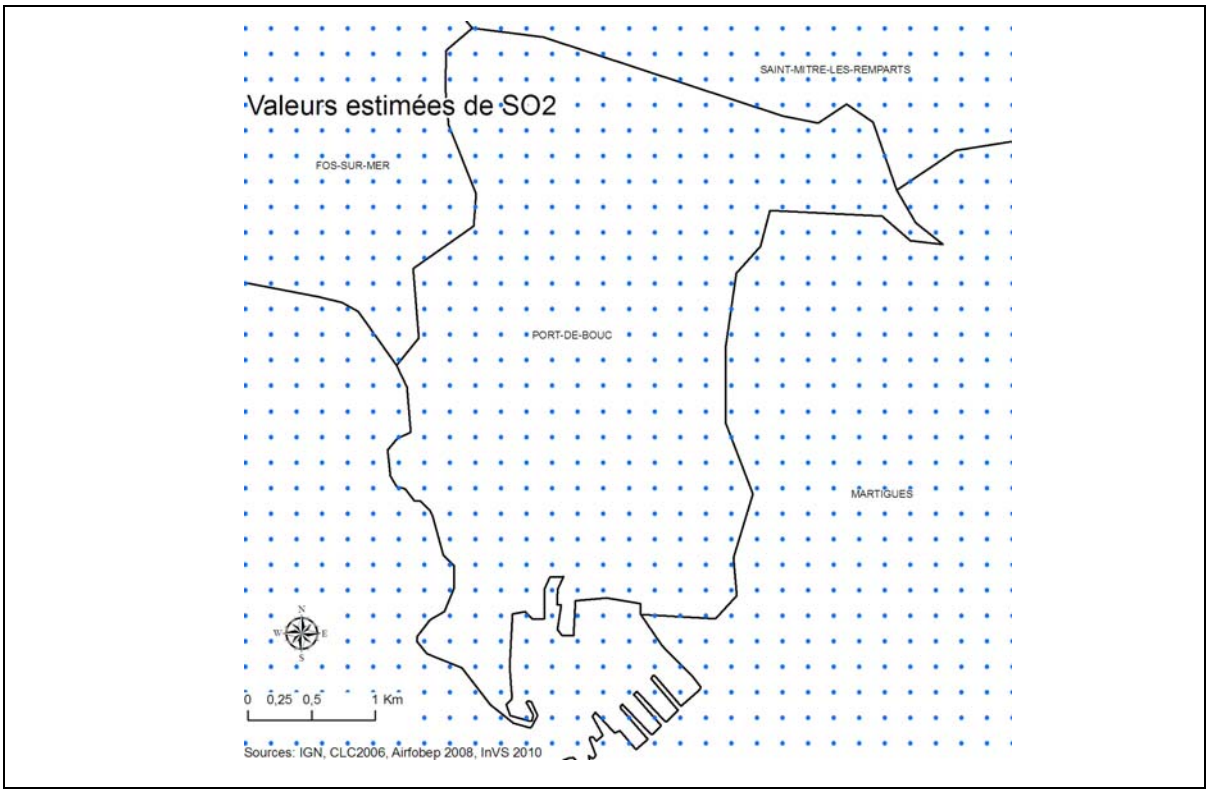
Dans une étude s'intéressant aux effets à long terme de la pollution atmosphérique sur les participants d'une cohorte (volet long terme du Programme de surveillance air et santé, cohorte de 20 000 agents d'Électricité de France - gaz de France), il était intéressant de regarder quelle est la proportion de sujets de la cohorte résidant à 2, 5, 10, 15 et 20 km d'une station de mesure de la qualité de l'air afin de juger de la pertinence d'utiliser les mesures de ces stations pour estimer l'exposition des sujets à la pollution atmosphérique. La sélection des codes postaux en fonction de leur distance aux stations a été possible dans le SIG grâce à l'outil de création des zones tampon.

Ces opérations peuvent être menées sur une ou plusieurs couches d'informations géo-référencées pour délimiter des zones de proximité (zones tampon) ou des zones d'influence (polygones de Thiessen[1]) à partir de la distance euclidienne (à vol d'oiseau). Parmi les opérations sur plusieurs couches, on retrouve notamment l'intersection géométrique mais aussi la jointure spatiale entre deux couches, c'est-à-dire qu'un élément géographique (une commune par exemple) va être enrichi des caractéristiques des entités spatiales avec lesquelles il est intersecté (les valeurs estimées d'un polluant) (figure 14).

[1] *Pour un ensemble de points répartis dans le plan, les polygones de Thiessen délimitent autour de chaque point la zone à l'intérieur de laquelle on est plus proche du point considéré que de tout autre point. En d'autres termes, on délimite ainsi la zone d'influence de chaque point d'un point de vue géométrique.*

| FIGURE 14 |

Créer un indicateur d'exposition par jointure spatiale



La jointure spatiale de la couche de points des valeurs de SO_2 avec celle des limites de communes permet ici de calculer une moyenne des valeurs par commune et ainsi, créer un indicateur d'exposition à la commune qui sera exploité dans l'analyse des effets de la pollution atmosphérique sur la santé.

Parmi les opérations d'analyse géographique, il existe également le comptage de points à l'intérieur d'un polygone (qui peut servir au calcul de prévalence par quartier après avoir géocodé des cas par exemple) ainsi que les opérateurs de proximité pour la sélection des plus proches voisins, par exemple, ou encore la fusion d'entités spatiales pour changer d'échelle (fusionner des Iris pour obtenir des communes). Les exemples de traitements géographiques rendus possibles par la mise en œuvre des SIG sont nombreux.

Une fois les données géoréférencées recueillies et intégrées dans le SIG (indicateurs sanitaires, environnementaux et sociodémographiques), il est possible d'en observer les interactions spatiales. Les études santé-environnement se prêtent bien à l'analyse de la dépendance spatiale, et à celle de la similitude de l'incidence entre unités spatiales proches géographiquement. Pour une maladie comme le cancer, la dépendance spatiale peut être attribuée à la répartition spatiale non aléatoire d'une exposition environnementale, d'un ou plusieurs facteurs de risque connus on inconnus de la maladie. La mise en évidence d'une structure particulière de la répartition spatiale implique souvent la reconnaissance de la similarité de la fréquence de la maladie dans des régions spatiales proches géographiquement [26]. Ceci revient à mesurer l'autocorrélation spatiale et donc, à évaluer l'intensité de la relation entre la proximité des lieux et leur degré de ressemblance (les objets proches se ressemblent plus que les objets éloignés) [27]. Deux types d'indicateurs peuvent être utilisés : la proximité spatiale (mesure du plus proche voisin, figure 15) et le degré de ressemblance (indice de Moran, figure 16).

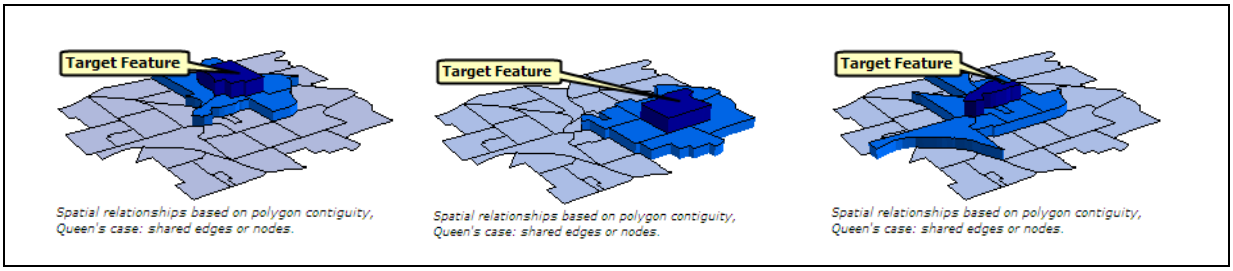
Les SIG permettent d'appréhender ces notions en complément de l'analyse statistique plus poussée. En effet, certains logiciels proposent de calculer des indices d'autocorrélation, des matrices de voisins, etc. et d'en cartographier les résultats.

C'est le cas du logiciel de SIG ArcGIS® qui est déployé à l'InVS. Ce logiciel dispose notamment d'outils qui permettent une première analyse descriptive des données géoréférencées à travers la représentation des résultats des calculs statistiques (figures 15, 16 et 16 bis).

Générer une matrice des voisins ou matrice de pondérations spatiales

| FIGURE 15 |

Copie d'écran de la construction d'une matrice des voisins dans ArcGIS®

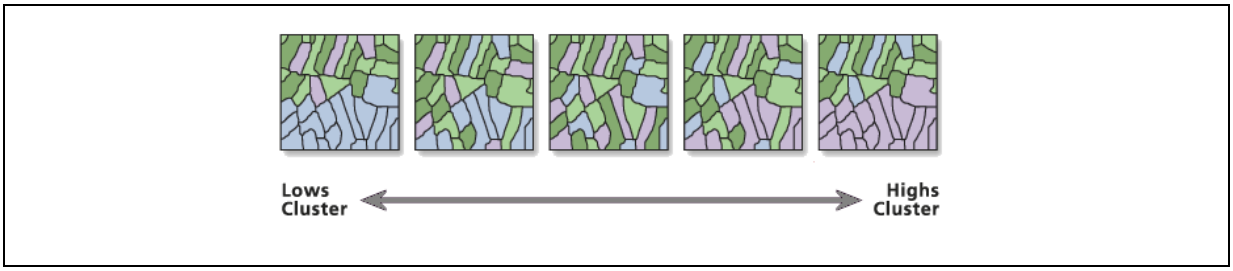


L'outil construit une matrice de voisinages afin de représenter les relations spatiales entre les entités d'un jeu de données. À chaque entité correspond un certain nombre d'entités voisines contigües comportant des caractéristiques spécifiques (indicateurs de santé, occupation du sol, etc.). Cet outil peut être utilisé pour délimiter une zone d'exposition par exemple.

Effectuer un test d'autocorrélation spatiale (indice de Moran I)

| FIGURE 16 |

Copie d'écran de la représentation de l'indice local de Moran dans ArcGIS®



L'indice local de Moran est un indice d'autocorrélation spatiale. L'outil que propose ArcGIS® permet de calculer cet indice et de représenter le résultat du test sous forme schématique (figure 16). À partir d'un ensemble d'entités pondérées, l'outil identifie l'emplacement où les valeurs élevées ou faibles sont regroupées dans l'espace, ainsi que les entités ayant des valeurs qui sont très différentes des valeurs d'entités environnantes.

Faire une analyse de concentration (statistique Gi* de Getis et Ord)

| FIGURE 16BIS |

Copie d'écran de la représentation de l'indice Gi* dans ArcGIS®

Rendered Gi* Output

L'analyse de concentration permet d'évaluer le degré de concentration géographique d'un semis de points dénombrés dans un maillage. L'outil calcule la statistique Gi*de Getis et Ord [28] pour l'analyse de semis de points, puis applique un type de rendu à tonalité froides-chaudes aux scores Z en sortie (score de concentration des ponctuels). La méthode suppose que la concentration dans une maille est indépendante de la concentration dans les mailles voisines [27].

Ces outils permettent une première description des données spatiales et ne sont qu'une étape dans la recherche d'agrégats spatiaux. Cette recherche est menée et approfondie grâce aux méthodes des statistiques spatiales dans un deuxième temps. Ces méthodes sont décrites et discutées dans le chapitre 3.

Grâce à l'installation de modules disponibles sous forme d'extensions, il est possible de disposer d'un certain nombre de fonctionnalités complémentaires, en sus des outils standards du logiciel, qui fonctionnent un peu comme des boîtes à outils spécialisés et adaptés à des problématiques plus spécifiques. Parmi elles, on peut évoquer l'extension Network Analyst qui permet le calcul de l'itinéraire optimal entre deux points à travers un réseau de rues avec la possibilité d'une réalisation en temps réel (exemple : aux États-Unis où de nombreuses villes disposent d'applications d'analyse de réseau dans un SIG afin de tracer le trajet le plus rapide entre un véhicule d'urgence et le lieu d'un accident). Cette extension a déjà été acquise temporairement à l'InVS à l'occasion du travail d'un stagiaire de Master 2 de géomatique sur le calcul d'accessibilité aux centres de soins pour les hémophiles. L'extension SpatialAnalyst que nous utilisons dans le cadre de nos travaux au DSE permet, quant à elle, en complément des outils d'analyse spatiale utilisés pour les traitements statistiques, de réaliser par exemple le calcul de densité de lignes ou de points, de mettre en œuvre des méthodes d'interpolation (voisin naturel, krigeage, spline, etc.), et d'effectuer des statistiques zonales, etc.

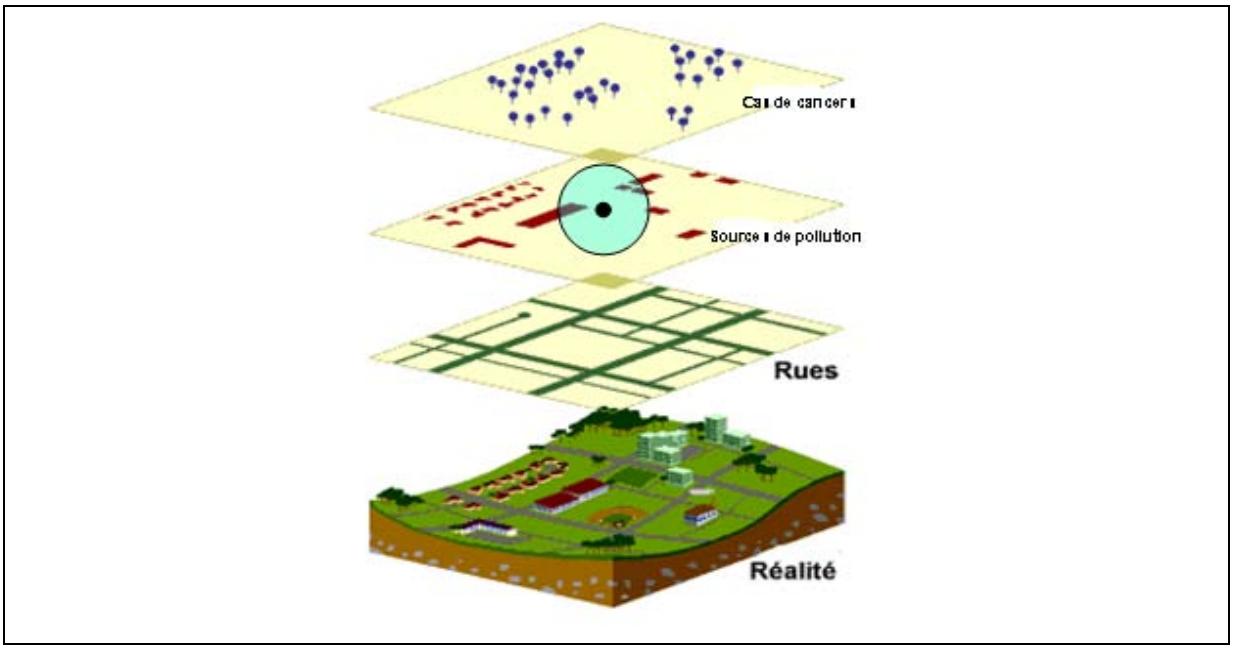
2.5 L'utilisation des SIG à diverses étapes d'une étude épidémiologique : l'exemple de travaux menés au département santé environnement de l'InVS

Les SIG sont sollicités à différentes étapes d'un projet (détermination de la composante spatiale d'une problématique, mise en place d'un protocole d'étude SIG, définition de la zone d'étude, construction d'indicateurs, production cartographique pour les rapports, etc.). Chaque étude, étant appliquée à un sujet spécifique, elle fait appel à des données spécifiques, à des unités spatiales définies, etc. Cependant, quelles que soient les thématiques et les problématiques d'étude, les méthodes et les étapes de la réflexion qui font appel à la mise en œuvre du SIG restent globalement toujours les mêmes.

En santé environnementale, comme nous l'avons dit, l'exploitation de la technologie associée aux SIG ne s'arrête pas à la représentation cartographique des indicateurs sanitaires. Les méthodes des sciences de l'information géographique constituent un réel atout lorsqu'il s'agit notamment d'appréhender les expositions à un facteur de risque environnemental donné. Les données disponibles pour estimer ces expositions ne sont pas toujours directement utilisables et intégrables dans un logiciel de SIG. Dans certaines situations, la donnée n'existe même pas en elle-même et le SIG est alors mobilisé afin de construire un proxy de l'exposition. Dans ce travail, la mobilisation de connaissances et de réflexions géographiques constituent un passage obligé.

| FIGURE 17 |

Superposition des données de natures différentes

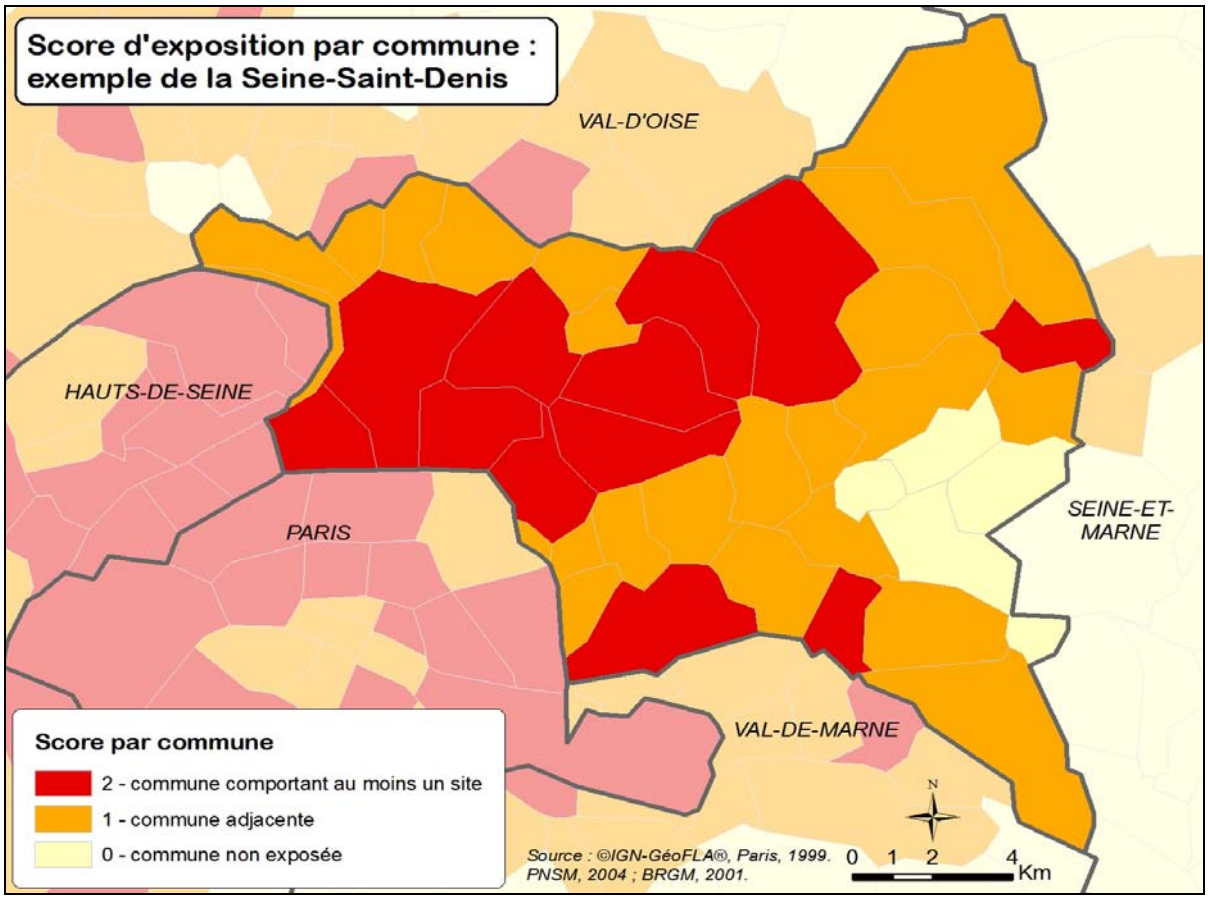


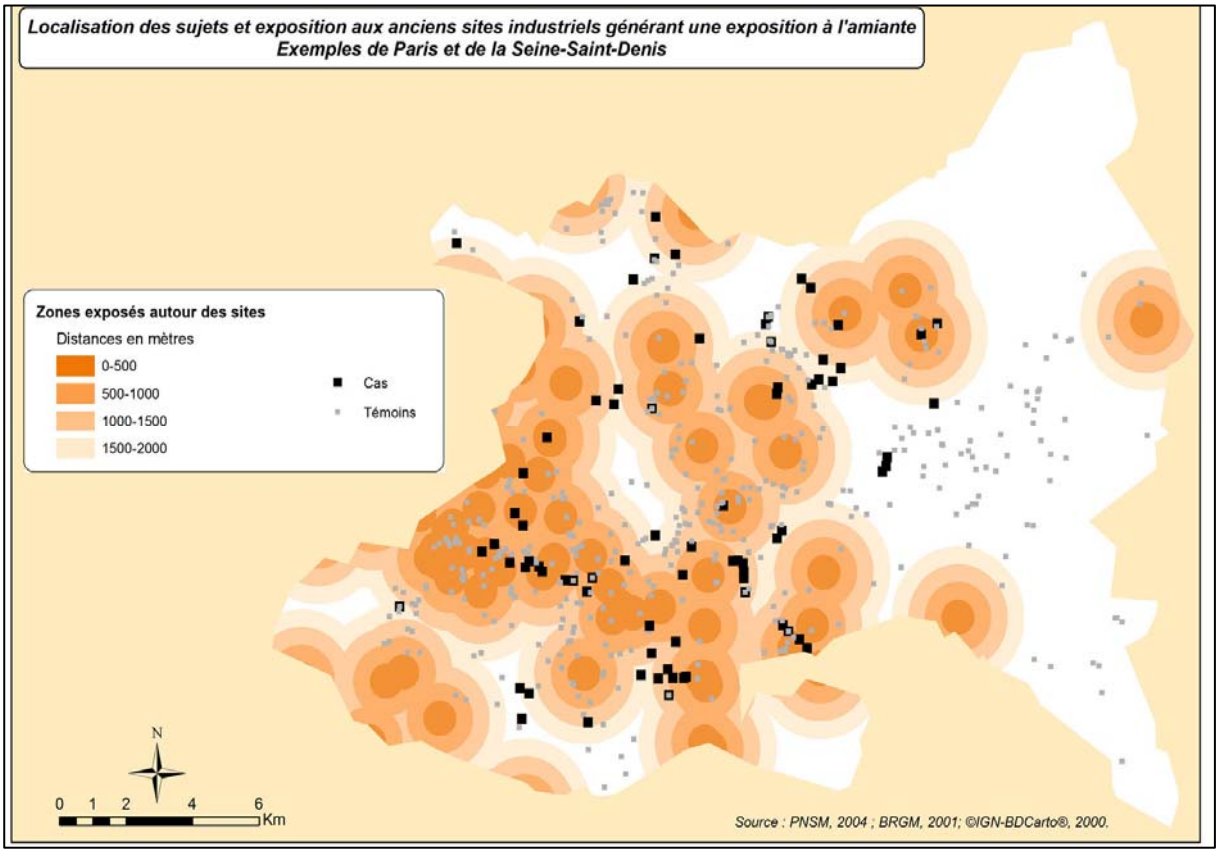
Les SIG sont particulièrement sollicités dans deux types d'étude principalement : les études locales autour d'un point source et les études écologiques géographiques, et c'est sur ces deux types d'étude que nous nous penchons ici plus spécifiquement. Les SIG sont en particulier mobilisés pour construire des indicateurs utilisés comme facteurs d'exposition ou d'exposition secondaire intégrés comme facteurs de confusion dans les analyses statistiques. Même si notre travail s'intéresse principalement aux études géographiques, il peut sembler intéressant de présenter rapidement ici les étapes de la construction d'un indicateur d'exposition dans une étude individuelle. Il s'agit de **l'étude cas-témoins sur l'exposition environnementale à l'amiante chez les personnes riveraines d'anciens sites industriels et affleurements naturels** [29]. À partir du positionnement des sites industriels susceptibles d'émettre de l'amiante dans l'environnement et des chantiers navals géoréférencés, le SIG a permis de construire un score d'exposition simple, reposant sur l'éloignement par rapport à ces sites. Avant de disposer d'un géocodage à l'adresse des sites, on a proposé un premier score simple qui se décline à la commune selon trois modalités (figure 18 - haut) : 2 – score le plus élevé : il se trouve dans la commune au moins un site rejetant de l'amiante dans l'environnement, 1 - la commune est adjacente (voisinage immédiat) d'une commune qui comporte un site rejetant de l'amiante et 0 – il n'y a pas de site rejetant de l'amiante dans la commune ou dans les communes immédiatement voisines. Suite à l'élaboration de ce premier score, un second, plus élaboré, a été construit après le géoréférencement à l'adresse des sites industriels et des individus de l'étude. Ce score, cartographié ici (figure 18 - bas), est fondé sur l'éloignement par rapport aux sites. Les distances prises en considération ont été choisies de manière à tenir compte le mieux possible des caractéristiques de dispersion des fibres d'amiante. Ce score, attribuant une valeur d'exposition à chaque adresse des sujets, a ensuite été exploité de façon plus complexe dans l'analyse statistique par la prise en compte de l'intensité de l'exposition selon les sites industriels sélectionnés, de la durée de l'exposition grâce à l'exploitation d'un calendrier résidentiel (figure 18) [29].

La construction d'indicateurs d'exposition et de confusion *via* le SIG intervient régulièrement, quel que soit le type d'études épidémiologiques. Le choix a été fait de présenter les utilisations des méthodes d'analyse mettant en œuvre des SIG à travers les études locales autour de points sources et les études de corrélation géographiques. Mais l'exemple décrit précédemment montre que ces méthodes sont tout aussi valables pour des études individuelles dès lors qu'elles comportent une forte composante spatiale.

| FIGURE 18 |

Étude cas témoins sur l'exposition environnementale à l'amiante





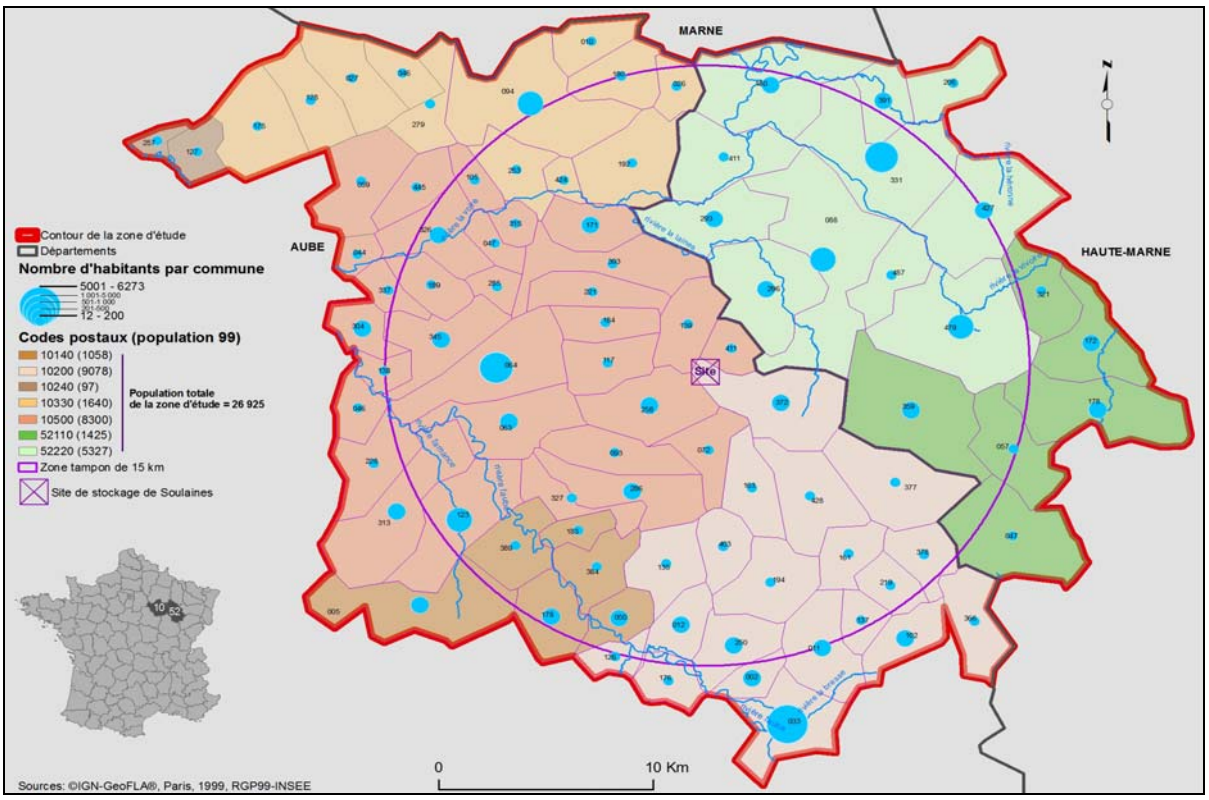
2.5.1 La mise en place d'un SIG dans une étude épidémiologique locale autour d'un point source

Dans ce type d'étude, la mise en place d'un SIG permet :

- la prise en compte de la distance au point source par la création de zones tampon dans une étude locale. Une méthode classique pour estimer l'exposition quand on a recours au SIG est d'utiliser la distance entre source et résidence de la population. Il s'agit de mesures de simple proximité qui ont tendance à surestimer la population réellement exposée mais c'est malgré tout un recours intéressant lorsque l'on ne dispose pas de données d'exposition plus précises (mesures, modélisations) [22] ;
- la sélection des communes concernées ;
- la création de variables nécessaires à l'étude :
  - par le calcul de la superficie de la zone d'étude,
  - par le calcul de la population totale concernée ;
- l'exploitation des résultats de la modélisation d'une exposition pour sa représentation cartographique (par la création de courbes d'iso concentration par exemple) et ensuite l'attribution des valeurs aux unités spatiales étudiées (figure 20).

| FIGURE 19 |

Étude en cours autour du site de stockage de Soulaines



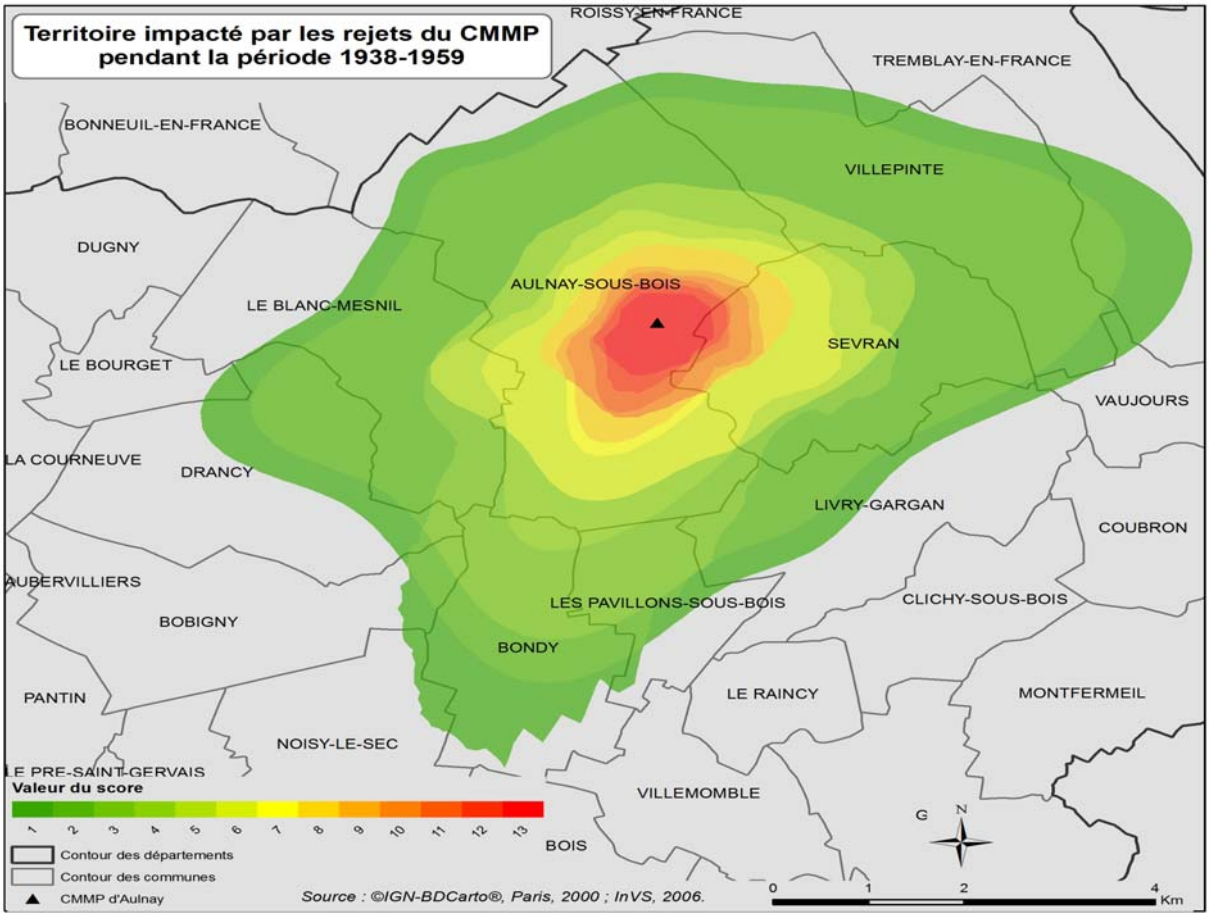
Le SIG permet d'introduire le facteur distance dans l'exposition à une pollution environnementale par la construction de cercles concentriques (appelés aussi buffers dans les logiciels de SIG) autour du point source et la sélection par entités spatiales des communes situées dans ce rayon donné. Dans l'étude de mortalité et d'incidence des cancers autour du centre de stockage de déchets radioactifs de faible et moyenne activité de Soulaines, illustrée ci-dessus (figure 19), la zone d'étude ressort en rouge ainsi que les communes dont le chef lieu est inclus dans un rayon de 15 km autour du site (cercle violet). On a croisé des données administratives (limites départementales et communales), démographiques (recensement de la population Insee), contextuelles (réseau hydrographique) avec le site de stockage afin d'avoir une vision d'ensemble de la zone d'étude, une estimation de la population potentiellement concernée, etc.

Le calcul de surface facilement calculable avec un SIG permet ensuite de calculer la densité de population, par exemple, mais aussi d'évaluer des dégâts et d'organiser les actions sur le terrain éventuellement.

Dans **l'étude de l'estimation rétrospective de l'exposition à l'amiante des populations avoisinantes du site de Comptoirs des minéraux et matières premières (CMMP) d'Aulnay-sous-Bois** [30], un SIG a été mis en place afin d'exploiter les résultats de la modélisation des rejets de l'usine réalisée grâce à un logiciel dédié (ADMS3®).

I FIGURE 20 I

Étude exposition à l'amiante autour du CMMP d'Aulnay-sous-bois



Cette étude, menée conjointement par la Cellule de l'InVS en région Ile-de-France et le DSE, comportait plusieurs volets dont un centré sur l'exposition aux rejets de fibres d'amiante dans l'environnement par cette usine pratiquant le broyage de matériaux, notamment de matériaux amiantés, ayant fonctionné de 1938 à 1975. Grâce aux archives et aux informations sur le fonctionnement de l'atelier, on a pu effectuer une série de modélisations de ces rejets en se basant sur différents scénarios et ce pour les deux périodes de fonctionnement différent de l'usine 1938-1959 et 1960-1975. On a distingué deux périodes d'étude car les modalités de ventilation et d'étanchéité des installations ont évolué entre ces deux moments du fait de travaux réalisés dans l'atelier.

Les résultats du modèle ont été cartographiés, après intégration des fichiers correspondant à des grilles de points de pas régulier au SIG, sous la forme de courbes d'iso concentration, à l'aide du module Spatial Analyst. La cartographie proposée utilise une sélection de courbes effectuée à la lumière des différents seuils de risques et des seuils réglementaires connus pour l'exposition à l'amiante.

Outre le fait que la visualisation apporte une lisibilité optimale d'un contexte, le SIG est en mesure d'aider à estimer la population impactée par les différents seuils cartographiés selon les différents scénarios. En partant du postulat, recevable puisque l'on se trouve en zone urbaine assez dense et que la répartition de la population se fait de manière homogène sur l'ensemble du territoire de chacune des communes concernées, la part de la superficie totale de chaque commune par les différentes courbes – correspondant aux différents seuils – est calculée. On en déduit l'effectif de la population impacté par ces différents seuils, en fonction des scénarios. Une cartographie de synthèse tous scénarios confondus est également proposée (figure 20). La limite de ce travail résidait principalement dans le fait que, d'une part, l'estimation des effectifs de population concernée était très tributaire du scénario modélisé, et d'autre part,

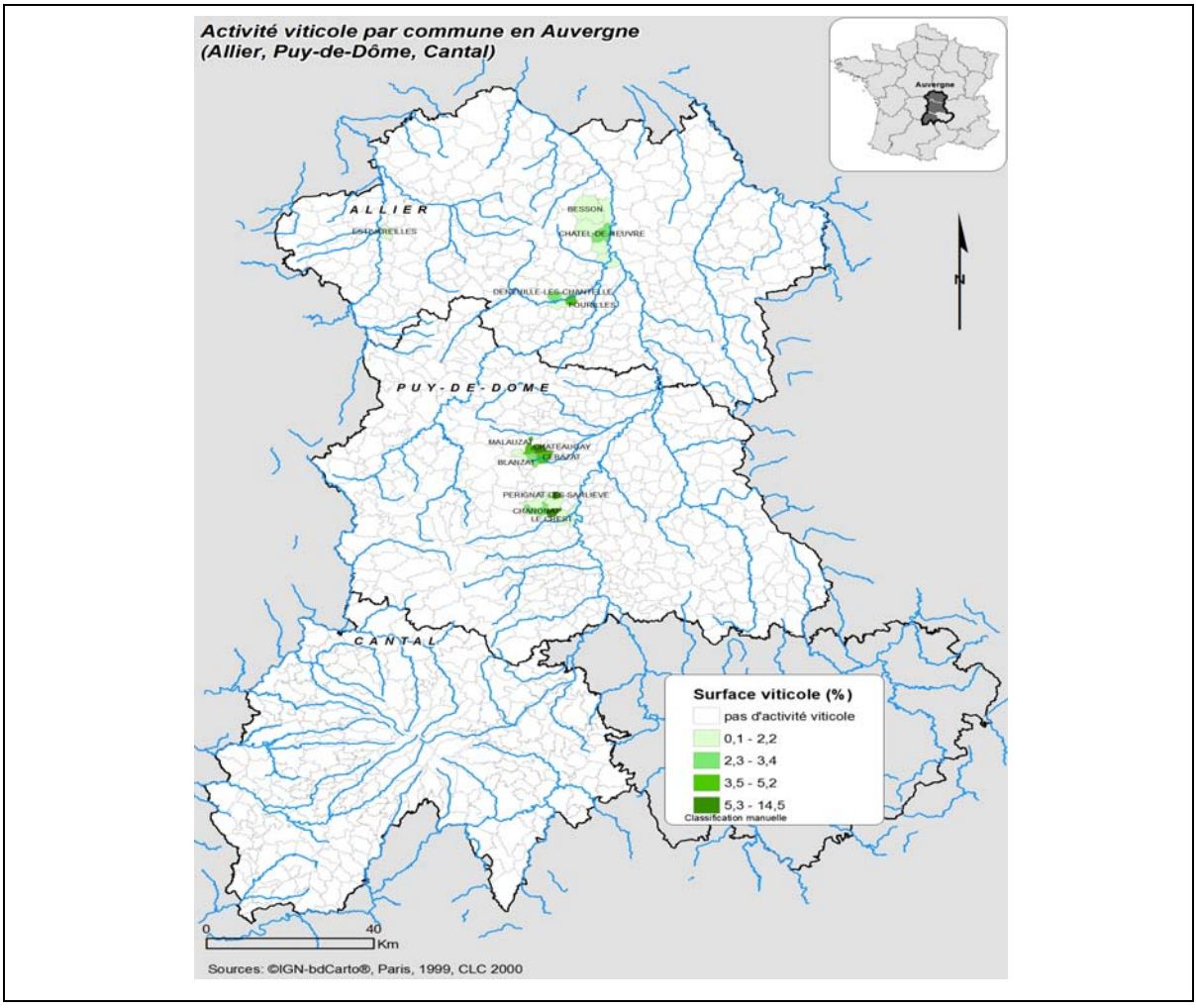
elle était très dépendante des seuils retenus pour la cartographie, ce qui peut poser problème dans la mesure où la relation entre l'exposition aux fibres d'amiante et la survenue d'un évènement sanitaire est considérée comme étant une relation sans seuil.

2.5.2 L'utilisation du SIG dans une étude de corrélation géographique

Dans ce type d'étude, mettre en place un SIG peut permettre de :
- déterminer l'unité spatiale de référence utilisée pour l'étude ;
- représenter des cas géocodés au préalable selon l'unité spatiale de référence choisie ;
- tenter une harmonisation des données n'ayant pas la même résolution afin de les rendre compatibles entre elles pour l'étude ;
- construire des indicateurs d'exposition et des facteurs de confusion[2]. On va alors combiner les données existantes et disponibles pour créer de nouvelles informations (indicateur d'exposition au trafic, indicateur d'exposition à une pollution de type industriel, etc.). Celles-ci sont ensuite reprises dans l'analyse statistique (détection de cluster, régression de Poisson, etc.).

I FIGURE 21 I

Étude arsenic hydrique en Auvergne



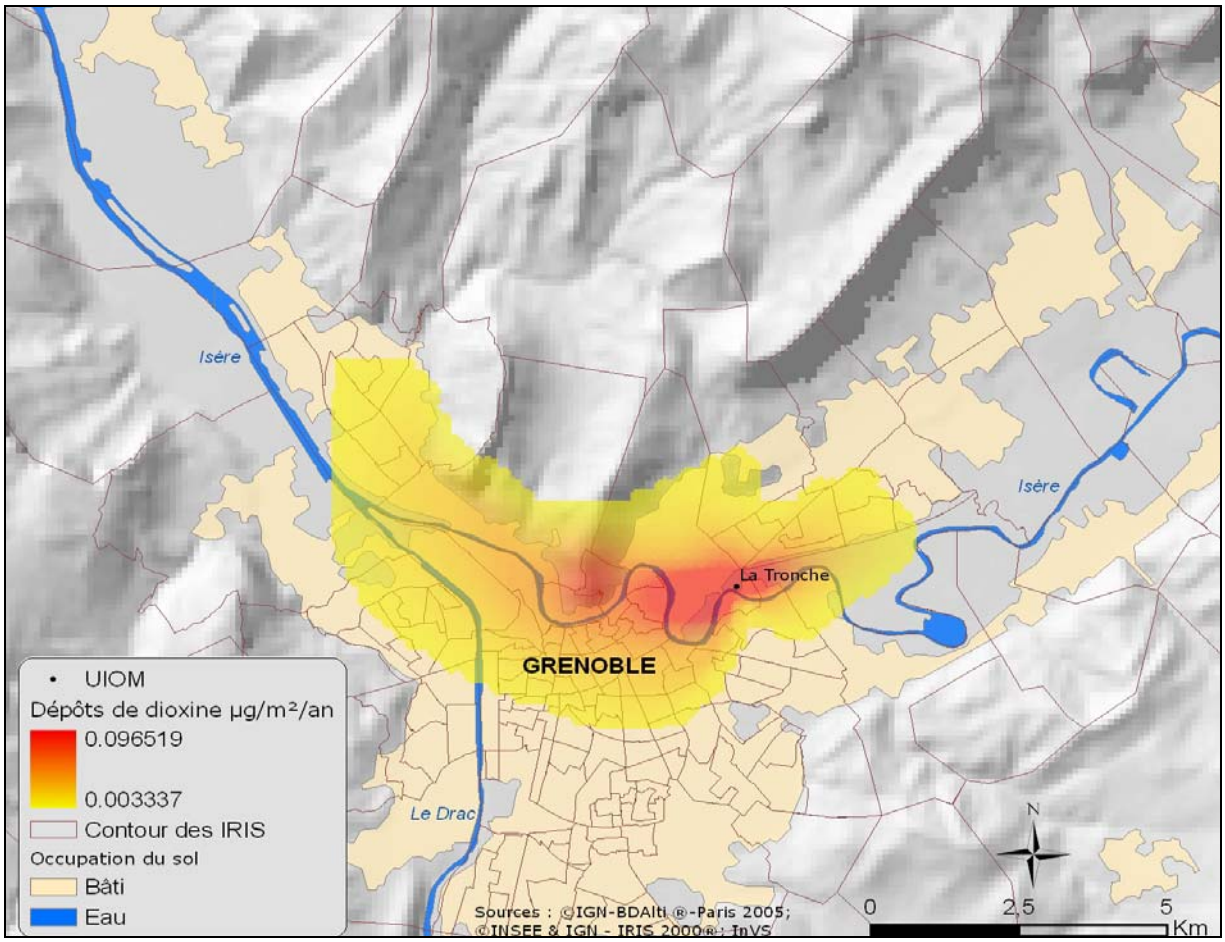
[2] *Facteur de confusion : facteurs d'exposition secondaires à prendre en compte afin de mesurer l'association entre l'exposition principale étudiée et les effets constatés sur la santé.*

Dans **l'étude sur l'arsenic hydrique et les cancers en Auvergne**, un certain nombre d'indicateurs d'exposition et de confusion ont été construits au moyen du SIG. Parmi eux, on peut citer l'exemple d'un proxy de l'exposition environnementale à l'arsenic présent dans les produits phytosanitaires utilisé dans la culture viticole. Le calcul de la surface viticole par commune (en pourcentage) a été retenu comme un bon indicateur de l'éventuelle pollution hydrique par les produits phytosanitaires pouvant contenir de l'arsenic et spécifiquement utilisés pour ce type d'activité agricole (figure 21). Ce nouvel indicateur est un facteur de confusion dans l'estimation de l'exposition des populations à l'arsenic hydrique et est utilisé dans l'analyse statistique menée dans un second temps.

Par ailleurs, dans **l'étude UIOM et cancers** [13] dont l'objectif était de déterminer si l'incidence des cancers est plus élevée chez les populations ayant été exposée aux rejets des incinérateurs d'ordures ménagères que dans la population non exposée, le SIG a été utilisé pour construire plusieurs variables. Cette étude écologique spatiale a mis en œuvre un nombre important de données de sources et de natures diverses et a impliqué la mise en place et l'exploitation d'un SIG complexe (coordonnées géographiques des 22 usines d'incinération d'ordures ménagères (UIOM), les cas de cancers fournis par quatre registres départementaux et géoréférencés à l'Iris, données d'altitude pour la modélisation des rejets atmosphériques, contours des Iris, données démographiques, etc.). C'est notamment grâce au SIG qu'ont été définis les départements d'étude et l'unité de référence spatiale qu'est l'Iris, qu'a été identifiée la population exposée après un gros travail de modélisation des rejets et d'exploitation de ces modélisations dans le SIG, qu'ont été construites les variables de confusions destinées à prendre en compte les spécificités géographiques des départements de l'étude. Les résultats de modélisations des émissions d'UIOM ont été récupérés sous la forme de fichiers textes correspondant à des grilles de points de 200 m de pas et de 20 à 40 km de côté centrées sur la cheminée de l'incinérateur et comportant, pour chacun de ces points de la grille, les coordonnées X et Y ainsi que des valeurs de concentrations et de dépôts modélisés. Par le croisement de ces panaches modélisés et intégrés au SIG avec les Iris, on a pu affecter à chaque unité géographique de l'étude une valeur de concentration et de dépôts (figure 22) [31].

I FIGURE 22 I

Étude UIOM



Cette étude constitue un bon exemple de l'utilisation des SIG pour la construction des indicateurs d'exposition. C'est ici le facteur d'exposition principal qui est décrit, mais le SIG a également été mobilisé pour la construction de facteurs d'exposition secondaires intervenant comme facteurs de confusion dans l'étude, notamment un facteur d'exposition au trafic routier.

La plupart des études menées jusqu'à aujourd'hui par le DSE impliquent une utilisation encore relativement simple des SIG (représentation des données géoréférencées, géotraitements et analyses assez élémentaires, croisement de diverses données de sources et de natures différentes, calcul de proximité, etc.). Pour autant, cette approche spatiale, à travers les SIG de plus en plus sollicités dans le milieu de la santé publique, offre des possibilités essentielles dans le domaine de la santé environnementale en permettant la visualisation et une meilleure connaissance du contexte géographique, environnemental et social d'une étude, la création d'indicateurs de pollution, le calcul de distance d'une population à un site pollué, etc.

2.6 Exemples d'utilisations des SIG en santé environnementale dans la littérature

La littérature montre que les études en épidémiologie environnementale utilisent très fréquemment les SIG pour divers aspects : le géocodage et la représentation des sujets d'une étude ou des sources de pollution, la définition de la population d'étude, l'identification des sources de pollution potentielles et des voies d'exposition, l'utilisation de la distance à un point source comme *proxy* de l'exposition pour évaluer l'exposition des personnes, l'intégration de données environnementales dans l'analyse d'un fait de santé.

L'objet de cette rapide présentation n'est absolument pas de donner une description exhaustive des utilisations des SIG au travers d'exemples donnés par la littérature d'études menées dans d'autres pays, mais plutôt, après avoir présenté plusieurs études menées au DSE de l'InVS, d'élargir le champ de la présentation. Voici deux brefs exemples d'utilisation des SIG en santé environnementale.

2.6.1 Croiser des données pour caractériser des populations potentiellement exposées

Exemple : Use of GIS and exposure modeling as tools in a study of cancer incidence in a population exposed to airborne dioxin [32].

Cette étude, réalisée au Danemark, a utilisé un modèle simulant l'exposition pour délimiter le plus précisément possible dans l'espace et dans le temps une population exposée aux dioxines dans l'air. Le SIG est utilisé pour faire le lien entre le modèle d'exposition et les données démographiques du recensement, les données individuelles (adresses, sexe, âge), les données de migration des sujets (déménagements dans, autour ou à l'extérieur de la zone) et les données de cancers. Le modèle a permis de découper la zone d'exposition en trois zones en fonction de l'intensité de l'exposition. En rendant possible la superposition des données individuelles, sanitaires et démographiques aux différentes zones d'exposition, le SIG a contribué à caractériser les populations et à décrire avec des méthodes statistiques complémentaires les liens santé environnement.

2.6.2 Création d'un indicateur d'exposition

Exemple : Residential exposure to petrochemicals and the risk of leukemia: using geographic information system tools to estimate individual-level residential exposure [33].

Dans une étude menée au sud de Taïwan sur l'exposition résidentielle aux sources de pollution pétrochimique et le risque de leucémies, Yu *et al.* ont utilisé un SIG pour estimer un niveau individuel d'exposition résidentielle. La mesure d'exposition attribuée au niveau individuel tient compte de la mobilité des sujets, du temps de résidence, de la distance aux installations pétrochimiques, de la direction du vent et de multiples sources de pollution pétrochimique. Dans le SIG, les auteurs de l'étude ont calculé la distance entre chaque lieu de résidence et les centroïdes des installations. La distance aux sites est pondérée par la direction des vents dominants. Les résidences situées dans un rayon de plus de 3 km autour des sites pollués ne sont pas considérées comme exposées. Le SIG a ainsi permis de construire un indicateur d'exposition à une pollution d'origine pétrochimique exploité ensuite pour la construction d'un modèle statistique pour l'analyse.

2.7 Conclusion et perspectives

L'apport des SIG, et plus globalement de la géographie en santé environnementale, n'est plus à démontrer. Cet apport est d'autant plus important si la réflexion géographique est intégrée au plus tôt dans le design de l'étude. En effet, le SIG est à considérer comme un outil de construction et de synthèse des différentes variables d'un modèle et comme base d'une analyse spatiale.

Mais malgré toutes les possibilités qu'ils offrent, décrites précédemment, il est indispensable de garder à l'esprit que les SIG et la cartographie ont leurs limites. Les attentes sont parfois trop grandes vis-à-vis des SIG qui sont parfois vus comme une solution technologique "miracle" permettant, par exemple, de surmonter certains obstacles rencontrés lors d'une étude.

- **Ce que les SIG ne permettent pas de faire :**
  - les données restent la plus importante des limites (disponibilité et qualité) et la mise en place d'un SIG dépend, comme le reste de l'étude, de leur disponibilité, de leur qualité, etc. ;
  - les SIG ne permettent pas de surmonter les limites des études écologiques géographiques qui imposent de considérer un groupe d'individus en supposant que tous ont les mêmes caractéristiques (socio-économiques, d'exposition, etc.) ;
  - les résultats issus d'un SIG et de travaux cartographiques doivent être considérés avec beaucoup de précaution *a fortiori* lorsqu'il s'agit d'études parfois sensibles ;
  - la cartographie ne représente souvent qu'un instant "t" alors qu'une exposition doit s'analyser dans la durée même si aujourd'hui des outils d'analyse spatio-temporelle (y compris parmi les SIG) commencent à se développer,
  - une mauvaise maîtrise des outils d'analyse spatiale peut entraîner une mauvaise interprétation des résultats ;
  - une gestion et une administration régulière du SIG sont indispensables pour en conserver la fiabilité (attention aux mises à jour). Par exemple, dans le cas d'une estimation des populations exposées à une pollution, il est vraisemblable que cette population évolue ou que de nouveaux sites apparaissent ;

- **Quelles perspectives pour les SIG au DSE ?**

Une nouvelle utilisation des SIG est développée sur 2010-2011 au DSE, dans le cadre du projet européen European Study of Cohorts for Air Pollution Effects. Il s'agit de la construction d'un modèle Land Use Regression [34], en France, pour l'estimation d'une exposition à la pollution atmosphérique développé pour la première fois dans l'étude Small Area Variation In Air pollution and Health (SAVIAH) pour modéliser les concentrations des NO_2 et basé sur les données géographiques locales [35]. Il s'agit d'un exemple d'utilisation de méthodes spatiales et des SIG dans des études épidémiologiques individuelles.

Les possibilités qu'offrent les SIG sont primordiales pour le développement de cette méthode. En effet, c'est dans le SIG que sont construites les variables explicatives autour des stations de mesures utilisées dans le modèle LUR : l'occupation du sol, le réseau routier, les données de trafic, les données topographiques et météorologiques (vents) et d'autres données encore (huit variables au maximum pour la lisibilité du modèle). Les variables en sortie sont ensuite exportées dans un logiciel statistique afin de construire le modèle de régression. Le modèle est alors utilisé pour estimer les concentrations de polluants au lieu de résidence de chaque personne incluse dans l'étude.

D'un point de vue méthodologique, le SIG intervient en amont et en aval de l'analyse statistique. Les deux démarches sont complémentaires, voire même souvent imbriquées : l'intégration des données spatialisées, la construction d'indicateurs et les premières analyses descriptives préparent l'analyse statistique qui permettra, dans une ultime étape, de cartographier les résultats pour la communication du rapport final (SIR lissés par exemple, incertitudes, résidus de modèles, etc.). Les deux disciplines ont une base commune : le choix de l'échelle, de l'unité spatiale de l'étude, des données géoréférencées disponibles.

3. Méthodes statistiques

Sont décrits ici quelques outils statistiques pour la détection de *clusters*, la représentation cartographique et les études de corrélation écologique utilisées au DSE de l'InVS. Pour une revue complète, il est indispensable de se référer aux ouvrages "Applied spatial statistics for public health data" de Waller et Gotway [36] et "Spatial epidemiology: methods and applications" de Elliott *et al.* [37] et au numéro de Statistics in Medicine, dédié aux méthodes de représentation cartographique [38].

Les outils statistiques utilisés en épidémiologie géographique ont connu un développement important depuis la fin des années 1980 et notamment, grâce au développement des techniques de Monte Carlo par chaînes de Markov. Le développement de ces outils était lié principalement à la nécessité de prendre en compte une possible surdispersion et autocorrélation spatiale qui n'étaient pas prises en compte par les modèles "classiques", et en particulier, par le modèle de Poisson. La surdispersion est définie par une variabilité du nombre de cas supérieure à celle attendue par la loi de Poisson. La présence de surdispersion peut témoigner d'agrégats (clusters) ou de la tendance des données à l'agrégation (clustering). L'autocorrélation spatiale est définie par la ressemblance des valeurs des taux d'incidence pour des zones voisines : le risque de maladie d'une zone géographique n'est pas indépendant de celui des zones voisines.

Dans les études de corrélation écologique, le contrôle des facteurs de confusion permet généralement de réduire la surdispersion et l'autocorrélation. Mais celles-ci peuvent être dues à des facteurs non mesurés ou à des erreurs dans les données qui ont une structure spatiale et il est alors important d'utiliser des modèles appropriés (modèles avec effets aléatoires, modèles bayésiens hiérarchiques) [14].

3.1 Détection de clusters et global clustering

De nombreuses méthodes ont été développées pour tester une tendance à l'agrégation de cas d'une pathologie [4]. Elles ont pour objectif de mieux comprendre la distribution géographique des maladies et d'en étudier l'hétérogénéité spatiale.

Une approche consiste à analyser globalement la distribution spatiale et temporelle d'une maladie. Une deuxième approche s'intéresse à l'estimation du risque d'une maladie par rapport à un point source.

Un cluster ou agrégat peut être défini comme une concentration de cas "anormalement élevée", supérieure à celle attendue, dans un groupe de personnes, une zone géographique ou une période de temps.

Les tests proposés dans le but de savoir si les événements sont agrégés dans l'espace peuvent être classés selon leur objectif.

De nombreux tests statistiques ont été développés pour étudier la variabilité spatiale d'une maladie, ceux-ci incluent les tests globaux pour évaluer la tendance globale au clustering ou à l'agrégation de l'incidence d'une maladie dans une région d'étude (les tests de corrélation spatiale, par exemple), les tests de détection pour identifier la localisation des clusters potentiels et tester si ces derniers sont significatifs et les tests focalisés ou de concentration utilisés quand une information permet *a priori* de spécifier une coordonnée géographique autour de la quelle la recherche d'un agrégat va se focaliser [36]. On présente et discute ces trois types de tests. Les méthodes de global clustering étudient la corrélation spatiale et détectent la tendance des cas à l'agrégation. Les méthodes de détection de cluster identifient les regroupements de cas incohérents avec l'hypothèse nulle de "no clustering" et évaluent leur niveau de significativité. La détection d'un cluster significatif n'implique pas une tendance globale au clustering significative et *vice versa* [39].

Les analyses de clusters peuvent être classées selon le type de données qu'elles permettent d'étudier [40]. Les deux catégories de données sont définies par leur niveau de résolution : elles sont soit agrégées ou de comptage (par exemple, le nombre de cas et la population par Iris ou commune de la zone géographique étudiée) soit ponctuelles ou individuelles (par exemple, les coordonnées spatiales des cas et de la population à risque ou des témoins). Nous nous intéressons ici aux données agrégées.

3.1.1 Détection de clusters et méthodes de balayage spatial

L'objectif des méthodes de balayage spatial est la surveillance géographique d'un territoire dans le but de détecter les zones pour lesquelles une incidence plus élevée de cas d'une maladie est observée, sans hypothèses *a priori*.

Les méthodes de balayage spatial cherchent à détecter l'emplacement des clusters dans la région étudiée. Elles appliquent des fenêtres (souvent des cercles) sur toute la région et dénombrent les cas et les individus à risque à l'intérieur et à l'extérieur de chaque fenêtre. Il existe différentes méthodes de balayage spatial, la méthode de Openshaw, la méthode de Besag et Newell et la statistique de scan spatiale [41], et elles se distinguent entre autres par la construction de la fenêtre qu'elles utilisent.

Méthodes de balayage spatial : la statistique de scan spatiale

Parmi les méthodes de détection de clusters, la statistique de scan spatiale [42-44] est devenue la plus populaire. L'objectif est d'identifier les zones ayant une incidence anormalement élevée et qui sont les moins "cohérentes" avec l'hypothèse nulle de risque constant. Cette méthode est basée sur un test du rapport de la vraisemblance. Cette méthode est très puissante et s'applique aussi bien sur des données groupées que ponctuelles.

Une fenêtre, de forme prédéfinie (cercles ou ellipses), de taille variable, balaye la zone d'étude. Pour chaque fenêtre, une statistique, basée sur le rapport de vraisemblance et les nombres de cas observés et attendus, est calculée. Les fonctions de vraisemblance s'écrivent selon le choix de la distribution théorique associée au nombre de cas. Deux distributions peuvent être définies : la loi de Poisson (données agrégées ou lorsque le nombre de cas est négligeable face à la taille de la population) et la loi binomiale (données individuelles des cas et témoins). L'hypothèse alternative, pour chaque "position spatiale" et taille de fenêtre, est qu'il existe un risque élevé à l'intérieur de la fenêtre par rapport à l'extérieur de la fenêtre. La fenêtre qui correspond au maximum de vraisemblance est le cluster le plus probable, celui qui a le moins de chance de survenir par hasard. Une valeur de p, calculée à partir de simulations de Monte Carlo, est assignée à ce cluster. La méthode de Kulldorff permet d'ordonner les clusters selon leur rapport de vraisemblance et identifier des clusters secondaires.

Le logiciel SaTScan® peut être utilisé pour mettre en œuvre la statistique de scan spatiale (et spatio-temporelle). Il s'agit d'un logiciel gratuit développé par Kulldorff [44,45]. SaTScan® permet de :
- détecter des clusters spatiaux ou spatio-temporels, et de voir s'ils sont statistiquement significatifs ;
- tester si la maladie est distribuée aléatoirement dans l'espace, le temps ou dans l'espace et le temps ;
- effectuer régulièrement la surveillance d'une maladie sur une zone géographique.

Le nombre de cas, la population et les coordonnées géographiques du centroïde (ou du chef-lieu) de chaque unité de la zone étudiée doivent être définis. Des covariables (sexe, classes d'âge, densité de population, score socio-économique…) peuvent être prises en compte. La taille du cluster maximal doit être définie et peut-être définie en fonction des effectifs de population. Souvent, dans la littérature, les clusters avec moins de 20 % de la population sont recherchés. On note que les clusters détectés ne peuvent pas être visualisés dans SaTScan®, pour cela on peut utiliser le package maptools du logiciel R [46] ou ArcVIEW®.

Les avantages de cette méthode sont :
- la prise en compte de covariables dans l'analyse ;
- la prise en compte des tests multiples - une valeur globale de p est fournie pour le test ;
- la localisation, même approximative, du cluster qui cause le rejet de l'hypothèse nulle, est donnée.

Les limites de cette méthode sont :
- les fenêtres sont des cercles ou des ellipses. La forme des agrégats potentiels doit être définie *a priori*. La partition spatiale de la région étudiée (et la partition temporelle de la période de temps étudiée) a une influence sur les clusters détectés. La statistique de balayage spatial tend à détecter des clusters de taille plus grande que celle des vrais clusters en englobant des régions voisines ou il n'y a pas de risque élevé [47] ;
- les frontières d'un cluster sont "incertaines". La localisation d'un cluster est "approximative".

D'autres méthodes de détection de clusters ont été développées notamment pour pouvoir détecter des clusters de forme arbitraire [47]. Mais, pour le moment, la méthode de balayage spatiale de Kulldorff est l'outil le plus utilisé pour identifier des clusters potentiels [39,48].

On insiste sur le fait qu'il est important de prendre en compte au moins la densité de population et éventuellement, un score socio-économique dans la recherche d'éventuels clusters.

3.1.2 Tests focalisés (ou tests de concentration)

De nombreuses méthodes permettent d'estimer le risque de maladie en relation à un point source. Ces méthodes ne s'intéressent pas à une tendance globale à l'agrégation mais à l'examen de l'existence d'agrégats en référence à un point spécifique.

Lorsque l'on dispose d'informations sur la position d'un "possible" cluster ou plutôt sur la position d'un point source, la statistique de balayage spatial ne doit pas être utilisée en raison d'une faible puissance induite par la prise en compte de toutes les localisations possibles alors que la localisation "supposée" est déjà connue.

Ces tests nécessitent une mesure du facteur de risque dans l'espace. Souvent, la distance au point source tient lieu d'indicateur d'exposition.

Il est important que le point source soit identifié en amont de la détection de clusters. Si au contraire, on commence par détecter le cluster le plus probable avant d'identifier le possible point source à proximité et que l'on calcule ensuite un test focalisé, alors l'hypothèse testée n'est plus la même et la valeur de p du test n'est pas correcte (page 252 de [36]).

Plusieurs tests sont disponibles [49] : le test de Stone du maximum de vraisemblance et le test du score de risque linéaire entre autres.

• Tests de Stone

Le test de Stone du rapport du maximum de vraisemblance et le test de Stone du maximum de Poisson sont utilisés pour tester une augmentation de risque en relation à un point source prédéfini [50,51].

Le test de Stone du rapport du maximum de vraisemblance est basé sur le rapport de vraisemblance. Il s'agit d'un test semi-paramétrique. Un index de rang, croissant avec la distance au point source, est calculé pour chaque unité géographique. Le nombre de cas observés est supposé indépendamment distribué selon une loi de Poisson. L'hypothèse nulle est l'égalité des risques dans les différentes unités géographiques. L'hypothèse alternative est la décroissance monotone du risque avec l'augmentation de la distance entre le point-source et les unités géographiques considérées (leur centroïde) ou plus précisément la décroissance monotone du risque avec l'augmentation des rangs de la distance entre le point-source et les unités géographiques considérées. Cette méthode est utilisée en général avec la distance mais peut être utilisée avec les rangs d'un indicateur d'exposition. La vraisemblance du modèle sous l'hypothèse alternative est comparée à celle sous l'hypothèse nulle. La significativité est examinée grâce à des méthodes de simulation.

Le test de Stone du maximum de Poisson définit une statistique de test égale à la valeur maximale observée du risque relatif (RR) obtenue en agrégeant les unités géographiques ordonnées par rapport à la distance du site en une zone de taille croissante. La significativité est examinée grâce à des méthodes de simulation.

Le test du rapport du maximum de vraisemblance est le plus utilisé des deux et semble être plus puissant [49].

Les tests de Stone sont très utilisés en épidémiologie et en particulier dans les études britanniques [49;52-54].

Les tests de Stone peuvent être utilisé pour tester l'augmentation de risque autour de plusieurs points source [52,54]. Mais ceci devrait être fait seulement si les points source sont comparables en termes d'exposition. Si une unité géographique est à proximité de plusieurs points source, une solution simple est de ne considérer que le point le plus proche [49].

L'avantage de ces tests demeure dans le fait de ne pas avoir à définir *a priori* la forme de la fonction de risque. En revanche, un point faible des tests de Stone est que la surdispersion n'est pas prise en compte. Les autres points faibles ou difficultés sont :

- le choix arbitraire de la distance maximale ;
- le choix arbitraire de la largeur des bandes autour du point source même si cette sélection est partiellement prise en compte dans les tests de Stone [52,54];
- le choix de la largeur des bandes quand plusieurs points source sont étudiés. Il est difficile de définir les bandes autour de plusieurs sites si, par exemple, certains se trouvent dans des communes rurales et d'autres dans des communes urbaines : pour certains points source et certaines distances, il pourrait ne pas y avoir de communes concernées.

Le package DCluster du logiciel R peut être utilisé pour calculer les tests de Stone [55].

• Test du score de risque linéaire

Le test proposé Bithell *et al.* [53] et Bithell [56] est un test basé sur le rapport de vraisemblance. Comme le test de Stone, il est utilisé pour tester une diminution du risque avec l'augmentation de la distance du site étudié.

Ce test alloue à chaque unité un score lié à une certaine mesure du risque (une mesure de l'exposition, la distance par rapport au point source ou un rang). Ces scores sont alors sommés pour toutes les unités géographiques pour obtenir un score global. Plusieurs scores peuvent être considérés. Bithell *et al.* [53] ont utilisé l'inverse de la distance et l'inverse du rang de la distance du centre de chaque unité étudiée par rapport au point source. Selon ces auteurs, l'inverse de la distance est approprié pour détecter un risque qui diminue avec la distance. Les rangs sont plus appropriés quand la proximité relative des cas au site est importante plutôt que la distance en elle même.

La significativité est examinée grâce à des méthodes de simulation.

Des études de puissance ont montré que deux tests du score de risque linéaire (en utilisant l'inverse de la distance et l'inverse du rang de la distance) étaient plus puissants que les tests de Stone [49].

Pour une présentation des tests basés sur la statistique du score de vraisemblance, nous pouvons nous baser sur la référence bibliographique [49].

Selon la référence utilisée, ces tests peuvent être conditionnels ou non conditionnels. Les tests conditionnels utilisent une référence interne à la zone étudiée : sous l'hypothèse nulle, les risques sont égaux à une constante inconnue ρ. Les tests non conditionnels utilisent une référence externe : sous l'hypothèse nulle, les risques sont égaux à 1. Les tests conditionnels considèrent seulement la distribution des cas dans la région étudiée et ils ignorent la différence entre le nombre de cas observés et le nombre de cas attendus autour du site. Les tests non conditionnels sont sensibles à un possible excès de risque dans la région étudiée comparés à une référence externe et à une possible distribution spatiale des cas observés.

"Ces tests focalisés ont une puissance faible pour détecter les petites augmentations de risque souvent associées aux exposition environnementales. D'où la nécessité d'utiliser plusieurs méthodes dans une même étude" [5].

3.1.3 Global clustering tests

Ces méthodes s'intéressent à l'existence d'une hétérogénéité globale de la distribution spatiale d'une maladie. L'objectif de ces méthodes est d'étudier la surdispersion et la corrélation spatiale et de détecter la tendance des cas "au clustering ". Ces méthodes ne donnent pas la localisation des clusters.

Il existe de nombreuses méthodes de global clustering, Kulldorff [57] en liste plus d'une centaine. On présente ici le test de Potthoff et Whittinghill, le test de Moran et le test de Tango qui sont très utilisés dans les études de corrélation spatiale.

- **Test de Potthoff et Whittinghill**

La première méthode consiste à tester l'existence d'une hétérogénéité spatiale globale en termes de surdispersion. Le test d'hétérogénéité le plus simple est celui de Pearson utilisant la loi du Chi2. Le test de surdispersion de Potthoff-Whittinghill est plus puissant dans le cas d'une hétérogénéité faible et il est largement utilisé en épidémiologie.

Sous l'hypothèse nulle d'une distribution aléatoire des cas d'une maladie, les taux d'incidence sont les mêmes sur toute la zone étudiée et les seules variations des cas observés sont liées aux fluctuations de la loi de Poisson. Le nombre de cas observés est supposé suivre une loi de Poisson de moyenne et de variance égale au nombre de cas attendus. Sous l'hypothèse alternative de l'existence d'une surdispersion des cas, un certain nombre de cas apparaissent dans certaines zones plus fréquemment que ce qui était prédit sous l'hypothèse d'une distribution de Poisson. Le rapport entre la variance et la moyenne du nombre de cas observés est supérieur à 1.

Le test de Potthoff et Whittinghill [41] suppose que le rapport entre la variance et la moyenne est égale à $1+\beta$, où β est défini comme la variation extra-poissonienne. Pour évaluer la surdispersion du risque de maladie, on évalue le rapport $\beta/ SE(\beta)$. En l'absence de surdispersion et lorsque le nombre de zones géographiques est grand, la distribution de $\beta/ SE(\beta)$ suit approximativement une loi Normale N(0,1).

Le package DCluster de R peut être utilisé pour calculer le test de Potthoff et Whittinghill [55].

- **La statistique de Moran**

Une deuxième méthode évalue l'existence d'une hétérogénéité spatiale globale en termes d'autocorrélation spatiale.

La statistique de Moran est l'indice d'autocorrélation spatiale le plus utilisé. Cette statistique résume le degré de ressemblance des unités géographiques voisines par une moyenne pondérée de la ressemblance entre observations. La statistique de Moran ne prend pas en compte l'hétérogénéité des effectifs de population : une corrélation spatiale significative pourrait être expliquée par la proximité de zones fortement peuplées et non pas par un cluster de taux élevés. Des versions alternatives de la statistique de Moran ont été proposées pour prendre en compte des effectifs de population hétérogènes [39].

Le package spdep de R peut être utilisé pour calculer cet indice [46].

- **La statistique de Tango**

Tango a proposé une statistique Excess Event Test pour l'évaluation du global clustering [58,59]. La méthode de Tango teste si les cas de maladie sont regroupés dans des clusters à l'intérieur de la région d'étude.

3.1.4 Conclusion sur la détection de clusters et le global clustering

Ces méthodes répondent aux objectifs suivants : tester si une maladie est distribuée aléatoirement dans la région étudiée ; détecter des zones à incidence élevée…

Huang *et al.* [39] comparent ces différents tests pour répondre aux questions suivantes : quelle méthode est la plus appropriée et/ou la plus puissante pour comprendre la distribution spatiale d'une maladie ? Est-il possible de fournir un guide pour l'utilisation de ces méthodes statistiques quand appliquées par exemple à des données de cancer ? Parmi les tests de global clustering considérés – Moran, Besag et Newell, Tango – le test de Tango semble le plus puissant. Parmi les méthodes de détection de cluster étudiées – la statistique de scan spatiale avec fenêtres circulaires et elliptiques et d'autres méthodes basées sur le rapport de vraisemblance… – la statistique de Kulldorff avec fenêtres elliptiques semble être la plus puissante.

Ces tests d'analyse de cluster ne peuvent être considérés que comme des méthodes de "dépistage", derrière lesquelles des études plus ciblées doivent être mises en œuvre pour confirmer (ou pas) les hypothèses qu'elles permettent de dégager [60]. Dans cette logique, l'utilisation de plusieurs tests, basés sur des hypothèses et des méthodes d'estimations différentes, paraît être une solution intéressante. De plus, la convergence/cohérence des résultats de ces différents tests devrait être recherchée.

3.2 Représentation cartographique des maladies (disease mapping)

La représentation cartographique des indicateurs de santé permet la description de leur distribution spatiale, la mise en évidence de zones avec un risque élevé pour la suggestion d'hypothèses étiologiques (caractéristiques partagées par les unités géographiques). La difficulté est de présenter des images fiables des variations géographiques des indicateurs de santé (séparer les réelles variations et le bruit inhérent, modéliser la structure de ces variations).

Les cartes de risque de maladies présentent souvent le SMR. Le SMR est défini par le rapport entre un nombre de cas observés et un nombre de cas attendus sous l'hypothèse d'une incidence de référence. Le SMR correspond à l'estimateur du maximum de vraisemblance du RR, les fluctuations aléatoires du nombre de cas de maladie observé étant modélisées par une loi de Poisson.

Mais, pour des maladies rares ou des petites unités géographiques, les SMR peuvent être instables et donner des excès de risque apparents.

Ce problème est du au fait de considérer les risques indépendamment, d'une unité géographique à l'autre, sans prendre en compte l'autocorrélation spatiale [6]. La corrélation ou dépendance implique que des zones proches géographiquement ont des risques similaires (facteurs de risque communs non mesurés).

3.2.1 Instabilité de l'estimateur de maximum de vraisemblance du RR

Soient Y_i le nombre observé de cas dans l'unité géographique i, E_i le nombre attendu de cas et θ_i le RR de l'unité i.

Pour le modèle : $Y_i \sim Poisson(E_i \theta_i)$

l'estimateur de MV de θ_i est : $\hat{\theta}_i = SMR_i = \frac{Y_i}{E_i}$

avec variance : $\mathrm{var}(\hat{\theta}_i) = \frac{Y_i}{E_i^2}$

On observe que les petites unités ou les unités avec des petits effectifs peuvent avoir une variance associée aux SMR très grande estimation du risque instable. La variabilité des SMR est différente selon les unités géographiques ce qui peut donner une représentation cartographique bruitée où les SMR les plus extrêmes correspondent le plus probablement aux unités les moins peuplées [26]. L'exemple suivant est pris de GeoBUGS [61] et illustre ce point.

Exemple : cancer de la lèvre en Écosse
Les taux de cancer de la lèvre dans 56 counties de l'Écosse pour la période 1975-1980 ont été analysés par [62], [63] et [14] entre autres.

Les SMR sont présentés dans la figure 23. Des figures 23, 24 et 25, on remarque que les valeurs les plus "extrêmes" sont basées sur un nombre faible de cas attendus.

Les cartes des probabilités d'excéder 1 sont encore moins informatives que les cartes des SMR : elles ne montrent pas les valeurs des risques, des "faibles" surincidences peuvent être mises en évidence pour des unités avec une forte population [14].

Des méthodes de lissage des SMR ont été développées pour produire des estimations plus fiables. L'intérêt du lissage est de permettre de mieux apprécier la structure spatiale sous-jacente en lissant le bruit causé par l'instabilité des SMR dans les zones à petit nombre de cas.

I FIGURE 23 I

Les SMR des 56 counties de l'Écosse

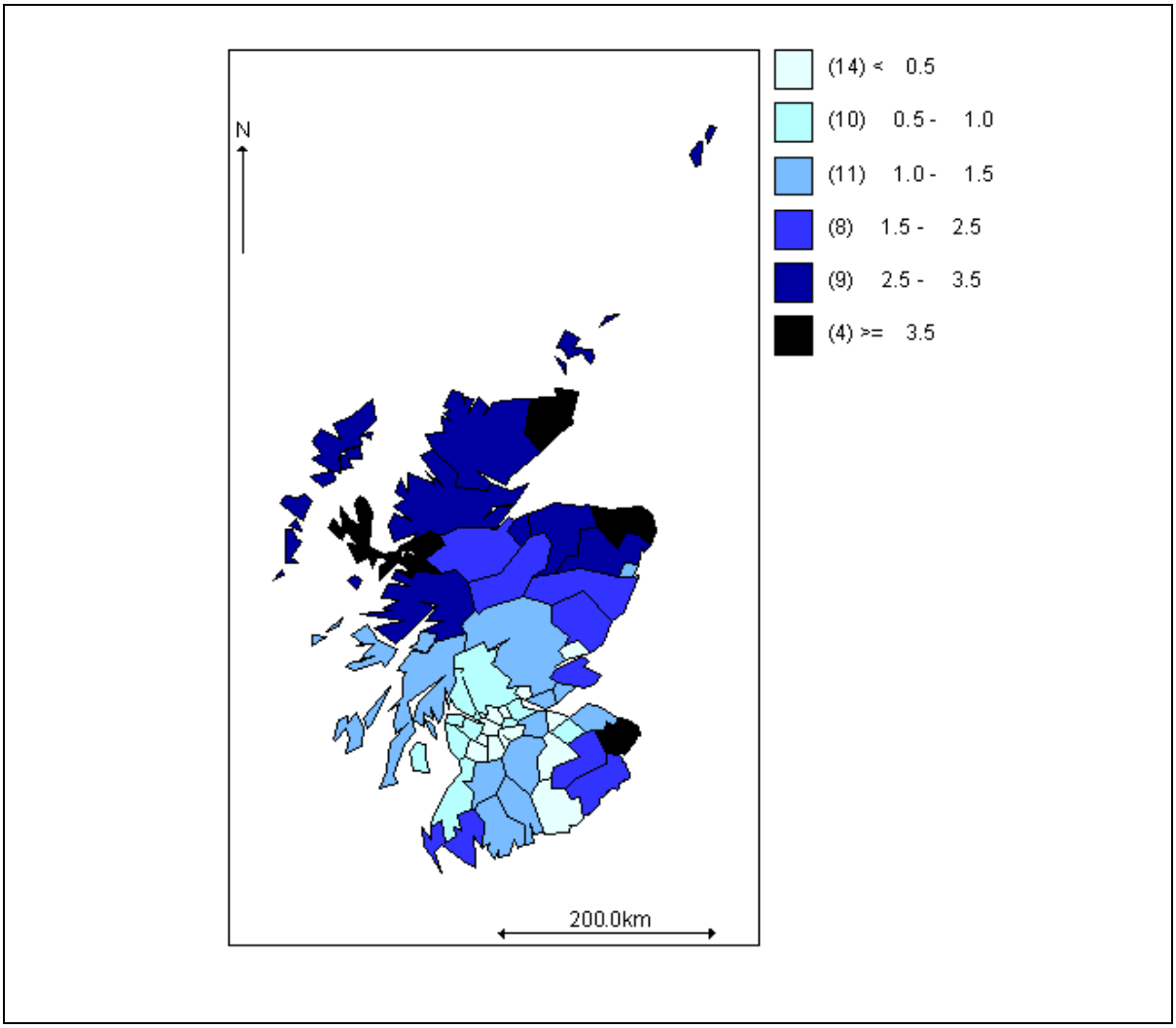


FIGURE 24

Le nombre de cas attendus varie entre 1.1 et 88.7

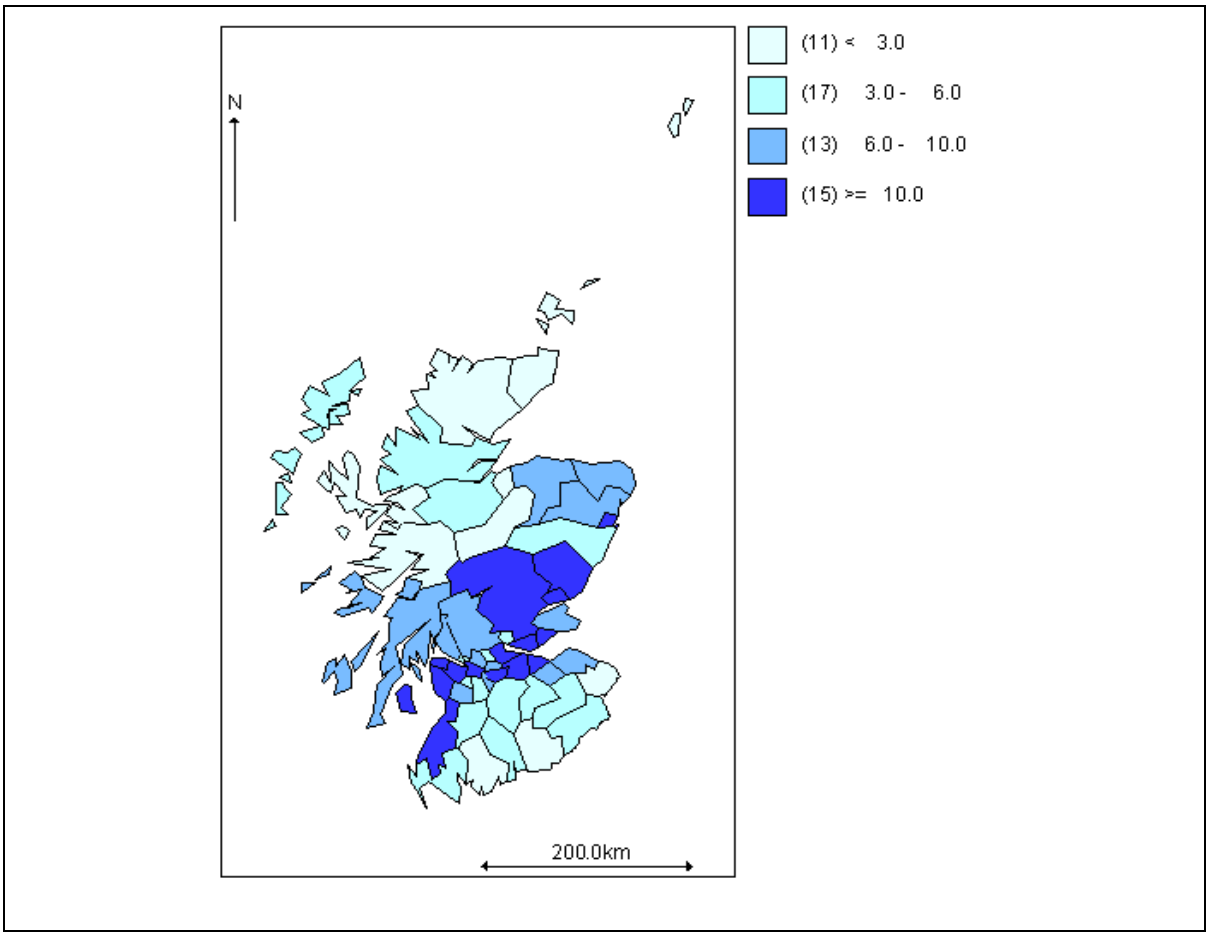
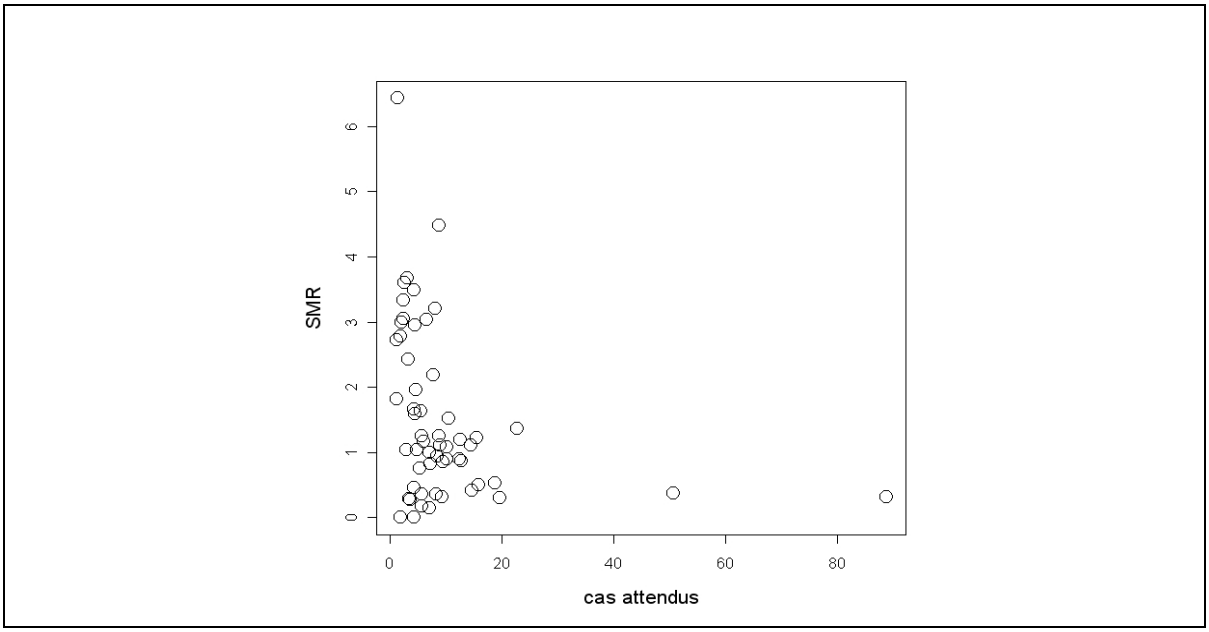


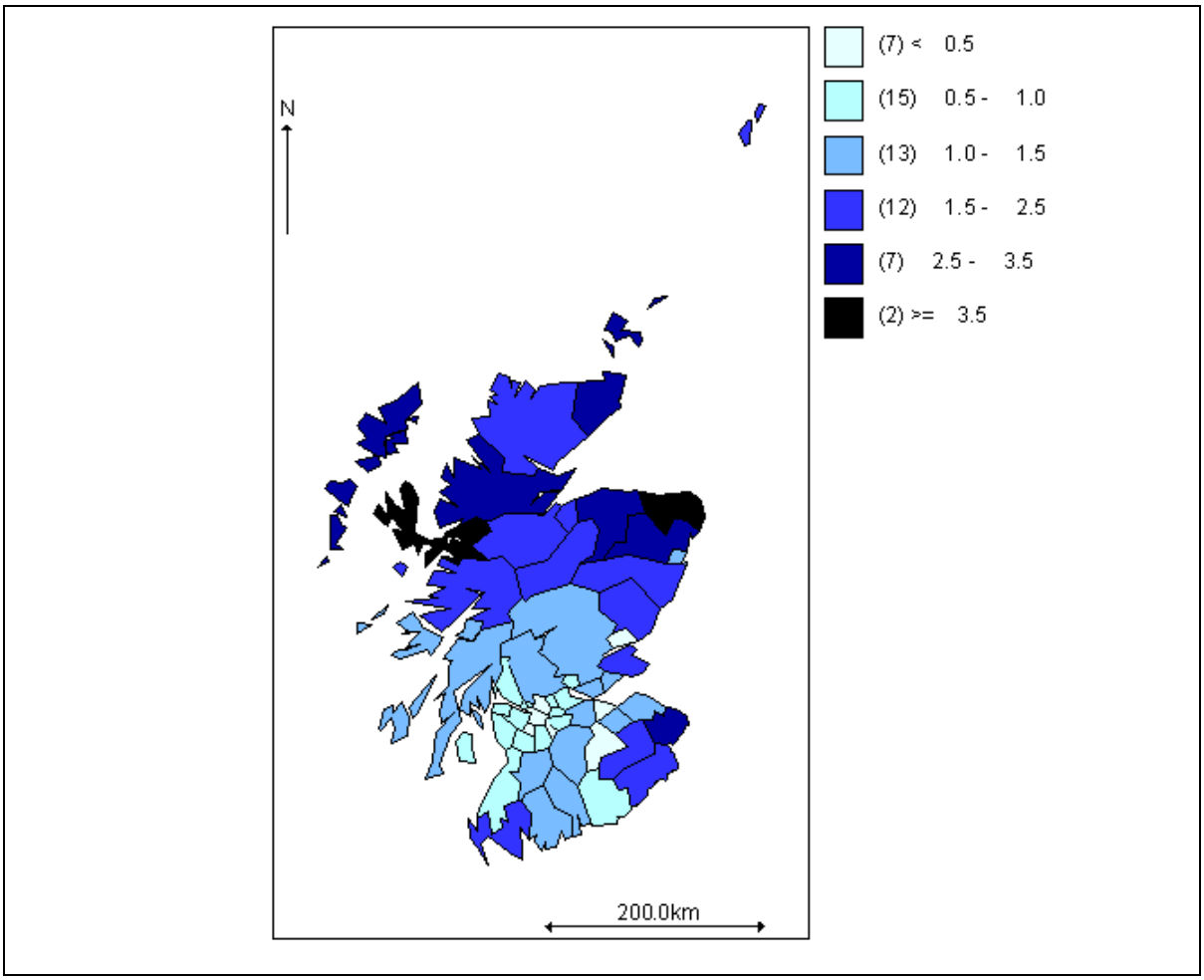
FIGURE 25

SMR et nombre de cas attendus



| FIGURE 26 |

Estimation bayésienne des RR des 56 counties de l'Ecosse (modèle Poisson-gamma)



3.2.2 L'approche bayésienne de lissage de taux

L'objectif est de lisser les différences de précision des estimations initiales, les SMR, en partageant l'information qu'apportent les différentes unités géographiques.

Les SMR peuvent être lissés en utilisant des modèles hiérarchiques qui prennent en compte les données de toutes les unités géographiques pour obtenir des estimations plus stables dans chaque unité géographique.

Dans l'approche classiquement utilisée, les observations de chacune des unités géographiques sont considérées comme une réalisation d'une variable aléatoire ayant une distribution de Poisson dont le paramètre, correspondant au RR, est considérée comme fixe et inconnu. Dans l'approche bayésienne, on suppose que ce paramètre est lui-même une variable aléatoire distribuant les risques entre les différentes unités, cette distribution étant appelée distribution *a priori*. L'estimation du RR est alors le résultat de la combinaison de l'information supposée *a priori* et l'information apportée par les observations. Lorsque l'information se situe essentiellement sur les données, la vraisemblance est importante. Lorsque les observations sont peu informatives, la connaissance apportée par la loi a *priori* devient primordiale. Un aspect essentiel de l'approche bayésienne concerne le choix de la loi *a priori* qui peut être déterminant dans les résultats [26].

Des modèles spatiaux ou des modèles non spatiaux peuvent être utilisés.

- **Modèles non spatiaux**

› *Modèle Poisson-Gamma*

Une première approche consiste en l'introduction d'une distribution *a priori* sur l'ensemble des risques. Cette structure globale sur tout le domaine empêche les estimations de prendre des valeurs trop grandes [6].

Soit :

$$Y_i \,|\theta_i \sim \text{Poisson}\left(E_i\,\theta_i\right)$$

Les risques relatifs, θ_i, sont supposés être indépendants et identiquement distribués selon la loi gamma :

$$\theta_i|\nu,\alpha \sim \text{Ga}\left(\nu,\alpha\right)$$

de moyenne ν/α et variance ν/α^2

La distribution de $Y_i|\nu,\alpha$ est binomiale négative. La moyenne et la variance de $Y_i|\nu,\alpha$ sont respectivement :

$$\text{E}\left[Y_i\,|\nu,\alpha\right] = E_i\,\frac{\nu}{\alpha},\ \text{Var}\left[Y_i\,|\nu,\alpha\right] = \text{E}\left[Y_i\,|\nu,\alpha\right]\left(1+\text{E}\left[Y_i\,|\nu,\alpha\right]/\nu\right).$$

Ce modèle est plus "raisonnable" que le modèle naïf de Poisson. Il prend en compte la dispersion extra-poissonnienne.

- **Estimation bayésienne empirique**

Si ν et α sont supposés connus la distribution *a posteriori* de θ_i suit une loi gamma. Si on a des estimations $\hat{\nu}$ et $\hat{\alpha}$, alors :

$$\theta_i|\mathbf{y},\hat{\nu},\hat{\alpha} \sim \text{Ga}\left(y_i+\hat{\nu},E_i+\hat{\alpha}\right).$$

et l'estimation *a posteriori* du RR de l'unité i est :

$$\text{E}\left[\theta_i|\mathbf{y},\hat{\nu},\hat{\alpha}\right] = \frac{y_i+\hat{\nu}}{E_i+\hat{\alpha}} = \text{SMR}_i\times\omega_i + \frac{\hat{\nu}}{\hat{\alpha}}\times(1-\omega_i)$$

une combinaison pondérée du SMR de l'unité i et de l'estimation *a priori*. Le poids associé au SMR de l'unité i est :

$$\omega_i = \frac{E_i}{E_i+\hat{\alpha}}.$$

On remarque que pour les unités avec une population importante l'estimation sera dominée par les données et sera proche du SMR. Pour les unités avec des effectifs faibles, le poids associé au SMR sera plus petit et le lissage sera plus important. Les estimations seront moins variables que les SMR (figure 26). Cette approche a donc pour effet d'atténuer les contrastes initiaux liés aux différences de précision des estimations.

Les estimations de ν et α de la binomiale négative peuvent être obtenues par MV [62].

- **Estimation complètement bayésienne**

Pour une approche bayésienne, une loi *a priori* est assignée aussi aux paramètres ν et α.

› *Modèle Poisson-lognormal avec effet aléatoire*

Comme le modèle Poisson-gamma, il s'agit d'un modèle bayésien hiérarchique caractérisé par :

- un premier niveau (variabilité locale d'événements rares) : la vraisemblance qui modélise la structure des observations. Le nombre observé de cas de cancer suit une distribution de Poisson :

$$(Y_i|\,\theta_i) \sim Poisson(E_i\,\theta_i)$$

- un deuxième niveau (structure interzones) : la distribution des risques relatifs. Ce niveau permet d'introduire la variabilité extra-Poisson :

$$\log(\theta_i) = \beta_0 + U_i$$

où β_0 est un terme constant qui représente l'effet moyen commun à toutes les unités géographiques et U_i sont des effets aléatoires gaussiens indépendants et identiquement distribués, $U_i \sim N(0, \sigma_u^2)$.

La définition des distributions *a priori* de β_0 et σ_u^2 est aussi nécessaire. La distribution marginale de ce modèle ne peut pas être calculée analytiquement. Il est nécessaire de faire appel à des méthodes de simulation (algorithme de Monte Carlo par chaînes de Markov). Le logiciel WinBUGS peut être utilisé [64].

Ce modèle est plus flexible que le modèle Poisson-gamma, il permet d'intégrer facilement des covariables et une structure spatiale entre les risques relatifs.

Modèles spatiaux

Une deuxième approche consiste à modéliser une structure de dépendance spatiale entre les risques relatifs. Les risques relatifs de chacune des unités sont dans ce cas influencés par les risques des unités voisines. On décrit ici le modèle proposé par Besag, York et Mollié [65] qui est le plus utilisé. Ce modèle partage le risque résiduel en un effet aléatoire non spatial et un effet spatialement structuré qui suit un modèle gaussien autorégressif conditionnel. Un autre modèle fréquemment utilisé est le modèle multivarié gaussien [36].

Une tendance à grande échelle, nord-sud par exemple, peut être prise en compte dans ce modèle (en incluant les coordonnées géographiques des centroïdes des unités spatiales).

› *Modèle BYM (Besag, York and Mollié)*

Le modèle hiérarchique bayésien de Besag, York et Mollié est caractérisé par :

- un premier niveau (variabilité locale ou intrazone) : la vraisemblance qui modélise la structure des observations. Le nombre observé de cas de cancer suit une distribution de Poisson :

$$(Y_i | \theta_i) \sim Poisson(E_i \theta_i)$$

- un deuxième niveau (structure inter zones) : la loi *a priori* des risques relatifs qui résume une information globale sur la similarité des risques θ_i, sur leur moyenne et leur variabilité. Ce niveau permet d'introduire la dépendance spatiale :

$$\log(\theta_i) = \beta_0 + U_i + V_i$$

avec U et V effets aléatoires décrivant l'hétérogénéité et la corrélation spatiale, respectivement (U et V indépendants). Les effets aléatoires U et V peuvent être considérés comme des variables latentes capturant les effets de facteurs de risque inconnus ou non mesurés non structurés spatialement et structurés spatialement, respectivement.

La composante d'hétérogénéité est supposée suivre une loi normale définie par :

$$U_i \sim N(0, \sigma_u^2)$$

où σ_u^2 contrôle la variabilité des RR, dans sa composante non spatiale.

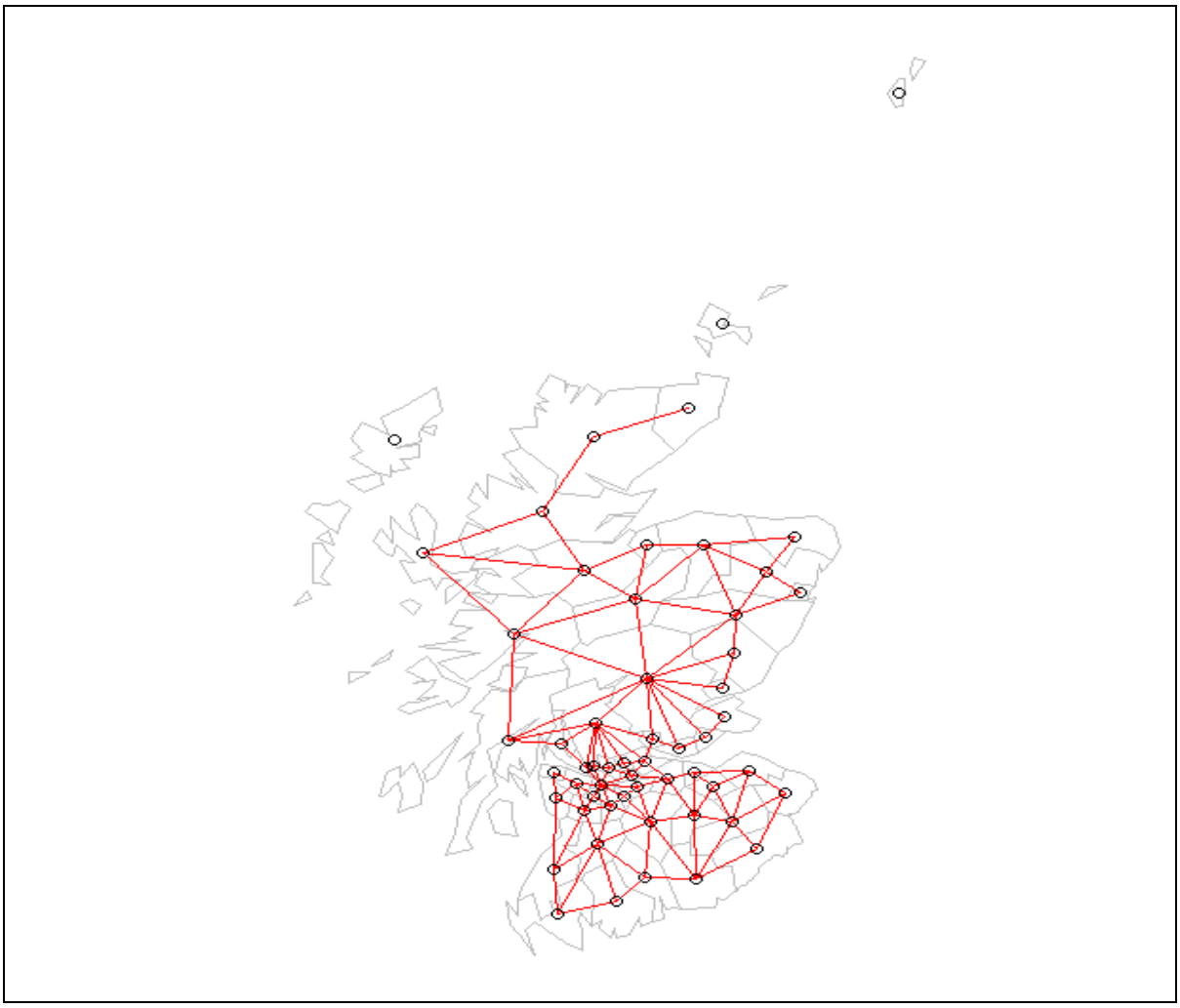
La composante spatiale suppose que les unités spatiales proches géographiquement tendent à avoir des RR similaires. Le modèle gaussien autorégressif conditionnel, modèle CAR intrinsèque, permet de prendre en compte cette hypothèse avec :

$$(V_i | V_j = v_j, j \neq i) \sim N\left(\frac{\sum_{j \neq i} w_{ij} v_j}{\sum_{j \neq i} w_{ij}}, \frac{\sigma_v^2}{\sum_{j \neq i} w_{ij}} \right)$$

où les poids w_{ij} décrivent la proximité géographique des unités i et j et σ_v^2 contrôle la variabilité conditionnelle des RR, dans sa composante spatiale. Le critère de proximité géographique le plus souvent retenu est celui d'adjacence. Les unités i et j sont voisines si elles partagent une frontière commune : $w_{ij} = 1$ si les unités i et j sont voisines et $w_{ij} = 0$ sinon (figure 27). Ce modèle suppose que la distribution conditionnelle de l'effet V_i dans l'unité géographique i est une loi normale centrée en la moyenne des effets de ses unités voisines et de variance inversement proportionnelle au nombre de voisins. Seul le paramètre σ_v^2 est libre.

| FIGURE 27 |

Exemple de la notion de voisinage selon le critère d'adjacence pour un modèle CAR



Le modèle CAR intrinsèque a l'avantage d'être facilement estimable. En effet, ses distributions conditionnelles complètes ont une forme analytique connue ce qui permet de recourir à l'échantillonneur de Gibbs. En revanche, ce modèle est impropre : sa moyenne est non définie et sa variance est infinie. La contrainte $\sum_i V_i = 0$ doit être imposée pour rendre le modèle identifiable.

Les variances σ_u^2 et σ_v^2 modulent les niveaux d'hétérogénéité globale et locale respectivement. Leur comparaison n'est pas immédiate car σ_v^2 est une variance conditionnelle qui dépend de la structure spatiale définie par les poids $\{w_{ij}\}$ alors que σ_u^2 est une variance marginale. Pour cela, il est utile de calculer aussi l'écart-type empirique de V_i, une estimation de la variabilité marginale des effets aléatoires spatialement structurés. Plus σ_u^2 est petit, plus les effets aléatoires ont tendance à être similaires entre toutes les unités géographiques. Plus σ_v^2 est petit, plus les effets aléatoires ont tendance à être similaires entre unités géographiques voisines. Il faut remarquer que, dans ce modèle, un seul paramètre, σ_v^2, contrôle la dépendance spatiale : si σ_v^2 est petit les résidus dépendent fortement de leur voisins mais la composante spatiale est "faible" dans le sens qu'elle contrôle peu la variabilité résiduelle [14]. Deux paramètres décrivent la dépendance spatiale dans le modèle multivarié gaussien.

Les distributions *a priori* de β_0, σ_v^2 et σ_u^2 doivent aussi être spécifiées. Le choix des distributions *a priori* des paramètres de variance est délicat [14].

Un gradient géographique, une variation lente et régulière à grande échelle, nord-sud par exemple, peut être aussi pris en compte dans ce modèle (en incluant les coordonnées géographiques des centroïdes des unités spatiales, par exemple).

Le résultat attendu est la loi *a posteriori* du risque de maladie. La distribution *a posteriori* est le produit de la distribution *a priori* et de la fonction de vraisemblance. Si les données sont informatives, la vraisemblance dominera la valeur estimée du RR ; dans le cas contraire, l'information apportée par la loi *a priori* aura un poids plus important. Les modèles hiérarchiques bayésiens permettent d'intégrer ces deux types d'information. L'estimation des paramètres de ce modèle fait appel à des intégrales qui ne sont pas calculables par des méthodes analytiques. Il est nécessaire de faire appel à des méthodes de simulation (algorithmes stochastiques de Monte Carlo par Chaînes de Markov).

Le modèle BYM a l'avantage de modéliser simultanément l'hétérogénéité globale et l'hétérogénéité locale des effets aléatoires. L'introduction de V_i permet de ne pas leur imposer la même variance pour chaque unité géographique puisque le nombre de voisins est différent pour chaque unité. Ce modèle donne un lissage mixte : un compromis entre lissage global (obtenu par un modèle Poisson-gamma, par exemple) et lissage local (obtenu par un modèle avec seulement la composante spatiale, par exemple).

Les modèles spatiaux posent le problème de choix de la structure spatiale des risques relatifs. La définition de voisinage est nécessaire pour le modèle BYM. Le voisinage le plus souvent utilisée suppose que deux unités spatiales sont voisines si elles partagent une frontière commune. D'autres voisinages peuvent être définis notamment à partir de la distance entre les centroïdes des unités de la zone d'étude.

Le modèle multivarié gaussien peut être aussi utilisé. Dans ce cas, la fonction de covariance doit être définie. Pour ce modèle, à partir d'une centaine d'unités spatiales, les temps de calcul peuvent être longs.

La faiblesse de ces deux modèles est liée au fait que les unités spatiales ne sont pas de forme régulière ou de population constante.

La mise en œuvre de ces modèles bayésiens est facilitée par l'utilisation du logiciel WinBUGS [64]. Ce logiciel repose sur l'estimation de la loi *a posteriori* par la méthode de simulation stochastique appelée échantillonneur de Gibbs. Cet algorithme permet de simuler un échantillon de la loi *a posteriori* jointe des paramètres du modèle. Si l'algorithme a convergé, c'est-à-dire simule correctement et suffisamment sous cette loi jointe, les inférences statistiques sont faites sur les lois *a posteriori* marginales de chacun des paramètres. Dans le cadre de nos travaux, nous nous intéressons particulièrement à la loi *a posteriori* marginale de chaque RR dont on retient la moyenne comme estimation bayésienne. Il est essentiel d'étudier la convergence de tels algorithmes.

- **Modèles spatio-temporels**

Les risques de maladie peuvent varier dans l'espace et dans le temps. Les modèles spatio-temporels sont utilisés pour décrire l'évolution dans le temps de la structure spatiale des maladies. Récemment, plusieurs modèles ont été proposés. Ces modèles peuvent être classés en trois catégories selon la structure d'évolution temporelle du risque de chaque zone : les modèles paramétriques (forme prédéfinie linéaire, quadratique…) [66], les modèles indépendants (les risques de chaque période sont estimés indépendamment des périodes précédentes) [67] et les modèles de lissage (ils permettent des tendances sans en prédéfinir la forme) [68-70]. L'évolution temporelle pour une zone géographique est déterminée par la somme de l'effet principal temps et des termes potentiels d'interaction qui incluent le temps.

Abellan *et al.* [71] montrent l'intérêt d'un modèle spatio-temporel pour l'analyse des malformations congénitales en Angleterre. L'inclusion de la composante temporelle permet d'étudier la stabilité de la distribution spatiale des maladies dans le temps. Ce qui renforce l'interprétation épidémiologique. Les auteurs soulignent qu'en effet deux situations très différentes peuvent donner le même nombre de cas "cumulé" dans le temps dans une zone géographique : a) un taux d'accumulation des cas constant dans le temps ce qui donne une distribution spatiale de la maladie étudiée constante dans le temps ou b) un taux d'accumulation qui varie fortement dans le temps et de manière différente pour certaines zones géographiques ce qui donne une distribution spatiale dans le temps particulièrement variable. Dans le premier cas, la distribution spatiale constante dans le temps pourrait être expliquée par des facteurs de risque constants dans le temps (sociodémographiques, environnementaux…). Dans le deuxième cas, elle pourrait être due à des facteurs de risque à courte latence qui pourraient créer des excès de cas dans des brefs intervalles de temps ou ces variations pourraient être dues à des changements "radicaux" d'enregistrement des cas.

Ugarte *et al.* [72] comparent différents modèles spatio-temporels bayésiens pour sélectionner ceux qui sont les plus adaptés à des données avec peu de fenêtres temporelles – en général, les données ne sont pas disponibles sur des longues périodes. Pour cela, les auteurs ont analysés les données de mortalité par cancer colorectal dans la région de Navarre en Espagne pour la période 1983-2002 (40 unités géographiques et quatre fenêtres temporelles de

cinq ans) et ils ont effectués des simulations pour analyser différents scénarios. La conclusion est qu'aucun modèle ne ressort comme étant le meilleur et que pour des analyses en routine le choix du modèle reste complexe.

L'estimation de ces modèles est faite en général *via* les algorithmes MCMC qui nécessitent un nombre élevé d'itérations afin de garantir la convergence indispensable à toute estimation. Les modèles spatio-temporels étant complexes les méthodes MCMC peuvent être très coûteuses en temps, modèles estimés en heures, voire en jours. De plus, le nombre de fois ou il est nécessaire d'utiliser les algorithmes MCMC peut être très important dans l'estimation de ces modèles. En effet, il est indispensable de faire des analyses de sensibilité aux différents paramètres des modèles (distribution *a priori*, par exemple). De nombreux travaux de recherche sont consacrés à accélérer et simplifier les algorithmes MCMC dans des modèles bayésiens complexes. Récemment, Rue *et al.* [73] ont développé INLA (Integrated Nested Laplace Approximations), un package de R, qui permet l'inférence bayésienne de modèles latents gaussiens. Cette méthode d'inférence repose sur des approximations de Laplace. La méthode développée permet une inférence bayésienne rapide, modèles estimés en quelques minutes, mais qui est limitée aux modèles latents gaussiens.

Plus de recherches sont nécessaires à cause de la complexité de ces modèles mais il s'agit d'un sujet de recherche très prometteur.

3.2.3 Conclusion sur les méthodes de disease mapping

Les méthodes présentées ont pour objectif de fournir des représentations cartographiques des risques qui soient le plus informatives possibles. L'intérêt du lissage est de permettre de mieux apprécier la structure spatiale sous-jacente en lissant le bruit causé par l'instabilité des SMR dans des unités à petit nombre de cas. L'enjeu de ces méthodes est de lisser les risques relatifs pour éliminer le bruit lié aux petits effectifs et en même temps, de ne pas trop lisser les risques relatifs pour pouvoir mettre en évidence leur structure spatiale. La distribution gaussienne utilisée dans le modèle CAR peut amener à un degré élevé de lissage. Des modèles alternatifs ont été développés pour permettre des éventuelles discontinuités, des changements abrupts dans la distribution spatiale des risques [74]. Le lissage conduit à réduire la sensibilité de la détection des unités à RR élevé. Pour remédier à cet inconvénient et augmenter cette sensibilité, Richardson *et al.* [75] proposent d'exploiter la distribution *a posteriori* des risques relatifs et définissent des règles de décision pour détecter les zones à risque élevé. Richardson *et al.* proposent de calculer à partir des résultats des simulations, la probabilité *a posteriori* que les risques relatifs soient supérieurs à 1 (avec une probabilité supérieure à 80 %).

L'objectif de ces études est de décrire la variabilité spatiale de la fréquence de la maladie. Elles permettent non seulement de mettre en évidence des contrastes entre les valeurs des indicateurs de santé mais aussi de suggérer et guider la recherche de facteurs de risque environnementaux pour formuler des hypothèses étiologiques. Ces études ont donc toute leur place dans le cadre de l'activité de veille sanitaire [7,14,26].

Dans les publications récentes, les principaux modèles développés sont axés autour de la description spatiale et spatio-temporelle des variations du risque d'une ou plusieurs maladies. Les modèles conjoints de plusieurs cancers sont développés pour rechercher des similitudes entres cancers (exposition environnementale commune) et aider à générer des hypothèses [76-79]. Les modèles conjoints de plusieurs maladies permettent de modéliser un effet spatialement structuré commun aux maladies étudiées. Cet effet aléatoire crée un lien de dépendance indirect entre les maladies étudiées et joue le rôle de substitut pour les facteurs d'exposition spatialement structurés mais non mesurés qui peuvent expliquer la répartition spatiale du risque de maladies. Il peut être intéressant d'analyser plusieurs maladies conjointement pour mettre en évidence des tendances de risque similaires qui pourraient refléter des facteurs de risque communs. De plus, une analyse multivariée peut donner une meilleure précision de l'estimation du risque d'une maladie en récoltant des informations (borrowing strength) sur d'autres maladies.

L'analyse de sources de données multiples se développe aussi dans les études épidémiologiques en vue notamment d'améliorer la fiabilité des diagnostics.

Les limites de ce type d'études sont liées à la faiblesse des effectifs, à la difficulté de prendre en compte le temps de latence suite à une exposition, à la nature de la maladie qui est le plus souvent multifactorielle, au problème d'évaluation de l'exposition (souvent de faible intensité et/ ou multiple), au problème des migrations surtout à un échelon géographique fin.

D'autres points importants pour la création de bonnes cartes concernent le choix de l'unité géographique, le choix de la méthode de discrétisation et le respect des règles de sémiologie graphique (chapitre 2).

3.3 Modèles de régression

L'objectif des analyses de régression dite "écologiques" est d'estimer l'association entre les variations géographiques d'un indicateur de santé et celles de variables environnementales [6].

Le fait d'étudier une maladie rare ou des petites unités spatiales conduit à utiliser un modèle de régression de Poisson. Les modèles sont les "mêmes" que ceux utilisés pour la représentation cartographique.

Mais il est important de souligner que les objectifs de la représentation cartographique et de la régression spatiale sont différents et la stratégie de modélisation doit refléter cette différence [14]. L'objectif de la représentation cartographique est la prédiction des risques relatifs par unité géographique alors que l'objectif de la régression écologique est l'estimation de la relation entre indicateur de santé et exposition.

Pour la présentation des modèles, il convient de se référer à la section "Représentation cartographique". Nous insistons ici sur quelques points qui nous paraissent importants.

On rappelle que le modèle de régression de Poisson classique est rarement adapté à cause de la sur-dispersion qui n'est pas prise en compte. Il est adapté quand la variabilité intrazone est négligeable comparée à la variabilité interzones (large zone d'étude et/ou maladies communes). Le modèle Poisson-lognormal avec un effet aléatoire capturant le log du RR résiduel/ inexpliqué peut être alors utilisé.

Il est peu réaliste de supposer l'indépendance des résidus de la régression : en général, les nombres de cas dans des zones voisines géographiquement présentent de la dépendance spatiale résiduelle. Dans le cadre de la représentation cartographique, cette dépendance peut être exploitée dans l'estimation des risques en lissant localement entre unités voisines. Dans le cadre de la régression, la dépendance doit être prise en compte et les méthodes statistiques classiques ne sont pas adaptées à l'analyse de données dépendantes. Le modèle BYM peut être utilisé. Il est important de vérifier la sensibilité des résultats à la structure spatiale considérée et aux distributions *a priori* des paramètres de variance.

Les modèles de régression écologiques posent le problème de choix de la structure spatiale des résidus. Différentes modélisations de l'autocorrélation des résidus existent dans la littérature. Pour une revue de la littérature, il convient de se reporter à l'ouvrage de Richardson [80]. Pour une comparaison de différents modèles spatiaux, c'est l'article de Best *et al.* Qui fait office de référence [81]. L'impact de la modélisation de la dépendance spatiale des résidus sur l'estimateur écologique doit toujours être étudié. Latouche *et al.* [82] ont étudié l'impact de la modélisation d'une sur-structure spatiale des résidus : le modèle imposait une structure spatiale alors que la variabilité spatiale de l'indicateur sanitaire était complètement expliquée par la variable d'exposition. Le modèle BYM était utilisé et ne semblait pas sous-estimer la relation écologique.

Lee and Durban [83] proposent le modèle "smooth-CAR" qui permet de séparer la tendance géographique à grande échelle et la corrélation spatiale locale.

Le choix d'introduire une tendance dans le modèle de régression n'est pas facile car l'exposition environnementale d'intérêt peut avoir aussi une structure spatiale. Si cette tendance peut être due à des facteurs de risque non mesurés alors elle doit être incluse dans le modèle [14].

Nous avons présenté l'approche bayésienne, une approche fréquentiste est possible aussi, mais peu de modèles ont été explorés. Un exemple est le "modèle additif binomial négatif" décrit par Thurston *et al.* [84]. Ce modèle permet de prendre en compte la surdispersion (modèle binomial négatif) et de modéliser la dépendance spatiale à grande échelle (modèle additif généralisé).

Ces modèles spatiaux peuvent dépendre fortement de l'unité utilisée. Modéliser les coordonnées géographiques des cas comme un processus ponctuel spatial est une approche alternative qui permet de ne plus devoir choisir une unité spatiale [85]. La difficulté est de définir une fonction d'intensité qui modélise la distribution de la population à risque.

Les variables écologiques sont souvent mesurées sur différentes échelles qui ne sont pas toujours emboîtées. En général, une transformation des données est effectuée pour les mettre toutes à la même échelle ce qui implique une perte importante d'information. Des modèles existent qui permettent de traiter des données de santé, de population et d'exposition disponibles à des échelles différentes [19].

L'inférence doit être faite au niveau agrégée de la zone étudiée, il est difficile de transposer les résultats au niveau individuel. Ce point est discuté dans le chapitre 1.

4. Un outil d'investigation rapide en santé environnement : The Rapid Inquiry Facility

Le Rapid Inquiry Facility (RIF) a été créé pour traiter rapidement des questions épidémiologiques et de santé publique. Il a été conçu par l'équipe du SAHSU du Département d'épidémiologie et santé publique de l'Imperial College de Londres (www.sahsu.org), afin d'analyser des données sanitaires en relation avec des expositions environnementales. C'est un outil qui permet de croiser des données sanitaires, environnementales, démographiques, de prendre en compte des facteurs de confusion et d'associer l'ensemble géographiquement.

RIF est basé sur un système d'information géographique et constitue une extension gratuite du logiciel de SIG ArcGIS® 9. Il permet aux utilisateurs un accès aux fonctionnalités qu'offre un SIG sans avoir besoin de connaissances approfondies du logiciel. Mais l'application étant intégrée dans ArcGIS®, il est possible d'utiliser toutes les fonctionnalités classiques du logiciel. Cet outil utilise à la fois l'approche base de données, la technologie des systèmes d'information géographique et l'approche statistiques spatiales.

RIF peut être utilisé pour l'analyse du risque sanitaire autour de sites polluants d'une part, et pour la représentation cartographique des maladies d'autre part. Il permet de calculer des taux standardisés d'incidence et des risques relatifs. Les données d'exposition (résultant de modèles de dispersion, par exemple) peuvent être importées et représentées dans le module cartographique d'ArcGIS®. De plus, il est possible d'exporter facilement les données vers SaTScan et WinBUGS pour mener des analyses complémentaires comme la détection d'agrégats ou le lissage spatial.

RIF est initialement conçu comme un outil pour l'équipe du SAHSU elle-même. Il a ensuite été adapté pour être utilisé par plusieurs pays européens dans le cadre des projets EUROHEIS et EUROHEIS2 (European Health and Environment Information System for Exposure and Disease Mapping and Risk Assessment).

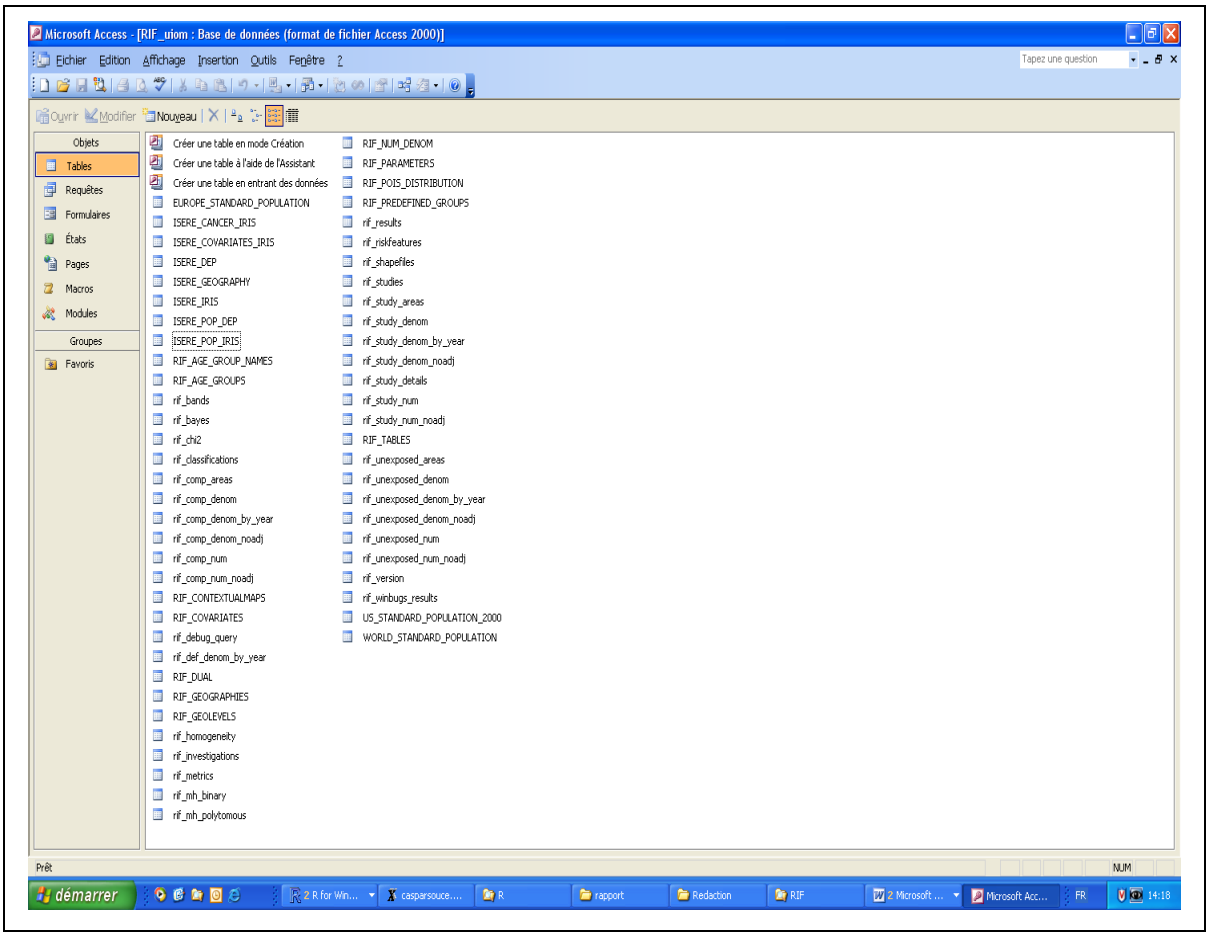
Les CDC (Centers for Disease Control and Prevention) et SAHSU collaborent également pour adapter RIF et pour pouvoir l'utiliser dans le cadre du CDC's National Environmental Public Health Tracking (EPHT) Network (réseau de surveillance santé environnement américain). L'objectif est de développer les fonctionnalités de RIF pour l'évaluation de relations spatio-temporelles entre maladies et expositions environnementales.

4.1 Méthodes de RIF

La première et principale étape à prévoir pour mettre en œuvre RIF est la construction de la base de données qui peut être faite sous Oracle ou Access. Elle doit respecter un format et une architecture spécifique et essentielle au fonctionnement de l'extension. Il est indispensable de définir certaines informations dans la base comme l'unité spatiale d'étude, la zone de comparaison pour le calcul des SMR notamment mais aussi, la population (par classe d'âge et sexe), les données de santé (cas de cancer par classe d'âge et sexe) et les covariables (niveau socio-économique, proximité aux sites pollués, etc.). Le bon fonctionnement de RIF dépend de la construction rigoureuse de la base de données (figure 28).

FIGURE 28

Exemple de base de données de RIF



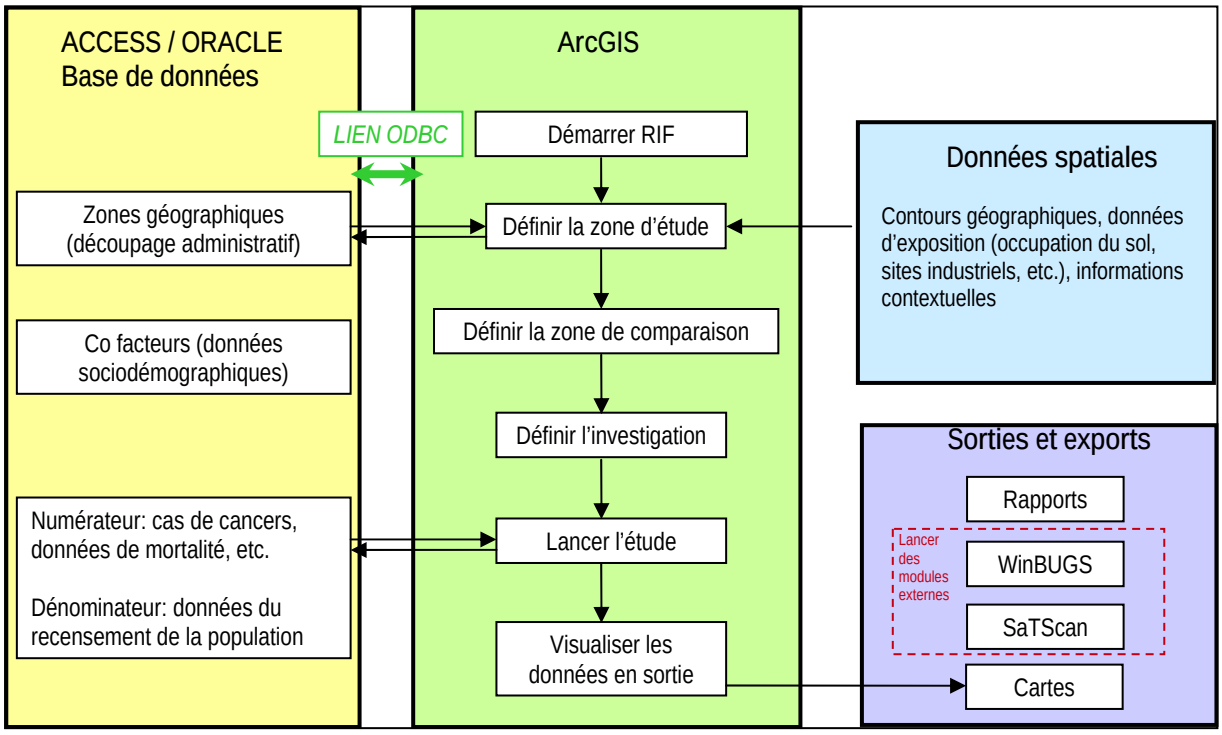
Cette base de données sera ensuite connectée au projet ArcMap© par un lien ODBC[3] (Open DataBase Connectivity) comme le montre la figure 29.

Le SIG à travers le logiciel ArcGIS© intervient en deuxième position une fois la base de données construite. Mais RIF est basé sur "la philosophie" SIG. Le logiciel de SIG est le support des développements de RIF qui en est une extension. Un simple onglet s'ajoute à l'interface habituelle. On a donc accès à toutes les fonctionnalités classiques d'ArcGIS© (ajout de couches, symbologie, mise en page, etc.). Le SIG est utilisé pour ses capacités de croisement de données, d'analyse et de communication (cartographie) et son utilisation est simplifiée au maximum afin d'être accessible aux non-connaisseurs du logiciel. Il permet de définir rapidement une population exposée, par rapport à son éloignement au site, en créant des cercles concentriques autour du point source d'une part ou par rapport aux valeurs d'exposition, si disponibles. Il permet la visualisation et l'analyse d'un contexte environnemental grâce à l'ajout d'informations sur l'occupation du sol, le réseau routier, etc.

[3] *Lien informatique vers une source de données à construire dans les outils d'administration du panneau de configuration du poste de travail.*

| FIGURE 29 |

Architecture de la base de données de RIF



Il est important de rappeler que RIF est axé sur l'investigation de problèmes à grande échelle (sur des petites unités géographiques comme la commune, l'Iris, l'îlot). En Grande-Bretagne, l'unité géographique utilisée est, par exemple, le zip code (code postal d'environ 10 000 personnes) ou encore le district (plus petite unité du recensement national, environ 400 personnes) et les régions pouvant être utilisées comme niveau de comparaison comptent autour de 10 millions de personnes.

RIF peut être utilisé pour deux types d'étude : l'analyse de risque autour d'un point source prédéfini et la cartographie des indicateurs de santé (RIF) Documentation: How to use the RIF ?).

L'objectif de l'analyse de risque autour d'un point source est de décrire le risque sanitaire à proximité du site étudié : observe-t-on un risque sanitaire plus élevé à proximité du site ? Pour cela, il faut avoir défini la zone potentiellement exposée et des éventuelles classes d'exposition. RIF permet alors de calculer les ratios standardisés (méthode de standardisation indirecte), SMR, leurs intervalles de confiance et de tester l'homogénéité de ces indicateurs. Pour cela, le test d'homogénéité du chi2 et le test de tendance linéaire sont réalisés. La figure 30 est une copie d'écran de RIF lors du calcul des SMR autour de plusieurs points sources (pour cet exemple, la distance est utilisée comme proxy de l'exposition et les SMR sont calculés par classe de distance).

La représentation cartographique des maladies permet de décrire la distribution géographique du risque d'une pathologie :
- le risque de maladie varie-t-il spatialement ?
- observe-t-on en moyenne ce que l'on attendait dans chacune des unités géographiques ?
- si des "écarts" sont observés, ont-ils une disposition géographique particulière – tendance à l'agrégation spatiale, gradient géographique ?

Pour cela, il est possible de produire les cartes des ratios standardisés. Le lissage bayésien des ratios standardisés est mis en œuvre (modèle Poisson-Gamma et modèle BYM) pour pouvoir prendre en compte l'instabilité de cet indicateur dans l'analyse de petites unités géographiques. La figure 31 est une impression d'écran de RIF où apparaît le module de représentation cartographique d'indicateurs sanitaires. On peut ainsi noter que l'outil détermine lui-même les bornes de la classification selon une méthode de discrétisation par défaut (méthode des seuils naturels) ainsi que la gamme de couleurs pour la représentation.

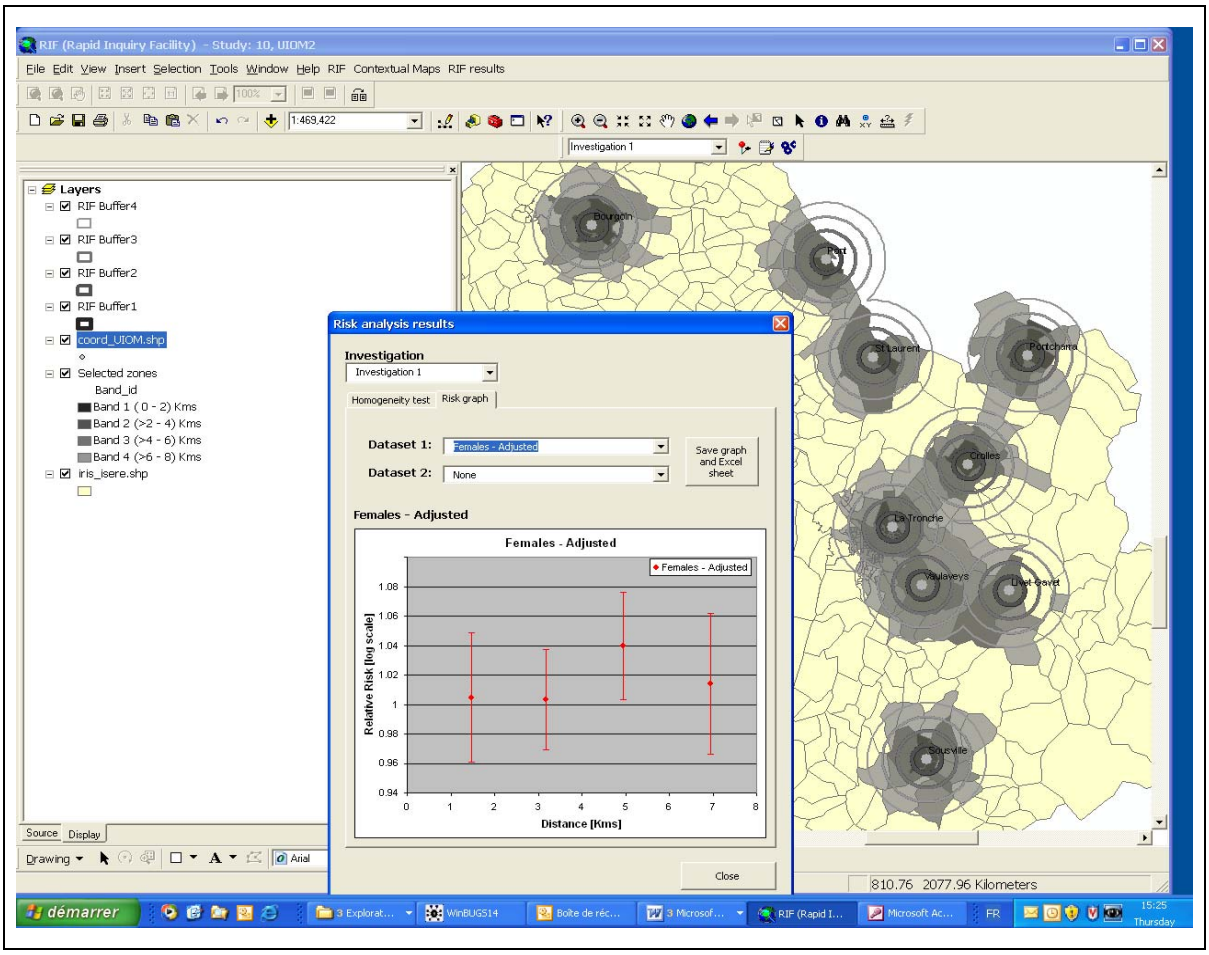
La démarche méthodologique de description des données dans RIF est parfaitement balisée et la plupart des outils sont déjà paramétrés, ce qui permet aux utilisateurs d'être guidés étape par étape. Ainsi, malgré le fait qu'il s'agisse d'une extension du logiciel ArcGIS®, RIF permet à un utilisateur non aguerri de ne pas se soucier de l'aspect purement cartographique. Il permet également d'exporter directement les cartes, graphiques et tableaux sous forme de rapports mis en forme de manière automatique. Il demeure possible de prendre la main sur la mise en forme, le choix des couleurs, etc., comme pour un projet SIG classique. De la même façon, les fonctionnalités de base du logiciel restant inchangées, les données géo-référencées sont très facilement ajoutées dans un projet RIF, comme par exemple, le résultat d'un modèle de dispersion qui pourra être importé et utilisé pour définir une population dans une analyse de risque. De la même façon, les informations contextuelles qui peuvent aider à l'interprétation peuvent aussi être affichées (réseau routier, occupation du sol, etc.).

Il s'agit ainsi d'un outil d'exploration descriptive rapide des données visant à avoir une première évaluation d'une situation. Mais RIF permet de creuser les investigations en rendant possible l'export rapide des données vers SaTScan® et WinBUGS® pour des analyses complémentaires et/ou plus "avancées".

Les utilisateurs doivent néanmoins avoir les mêmes considérations que lors d'une étude épidémiologique classique. Il reste à l'utilisateur le choix des pathologies d'intérêt, le choix des données (cohérentes avec la question posée), le choix de la population de référence, le choix de l'échelle géographique la plus appropriée qui dépend du contexte local (densité de population).

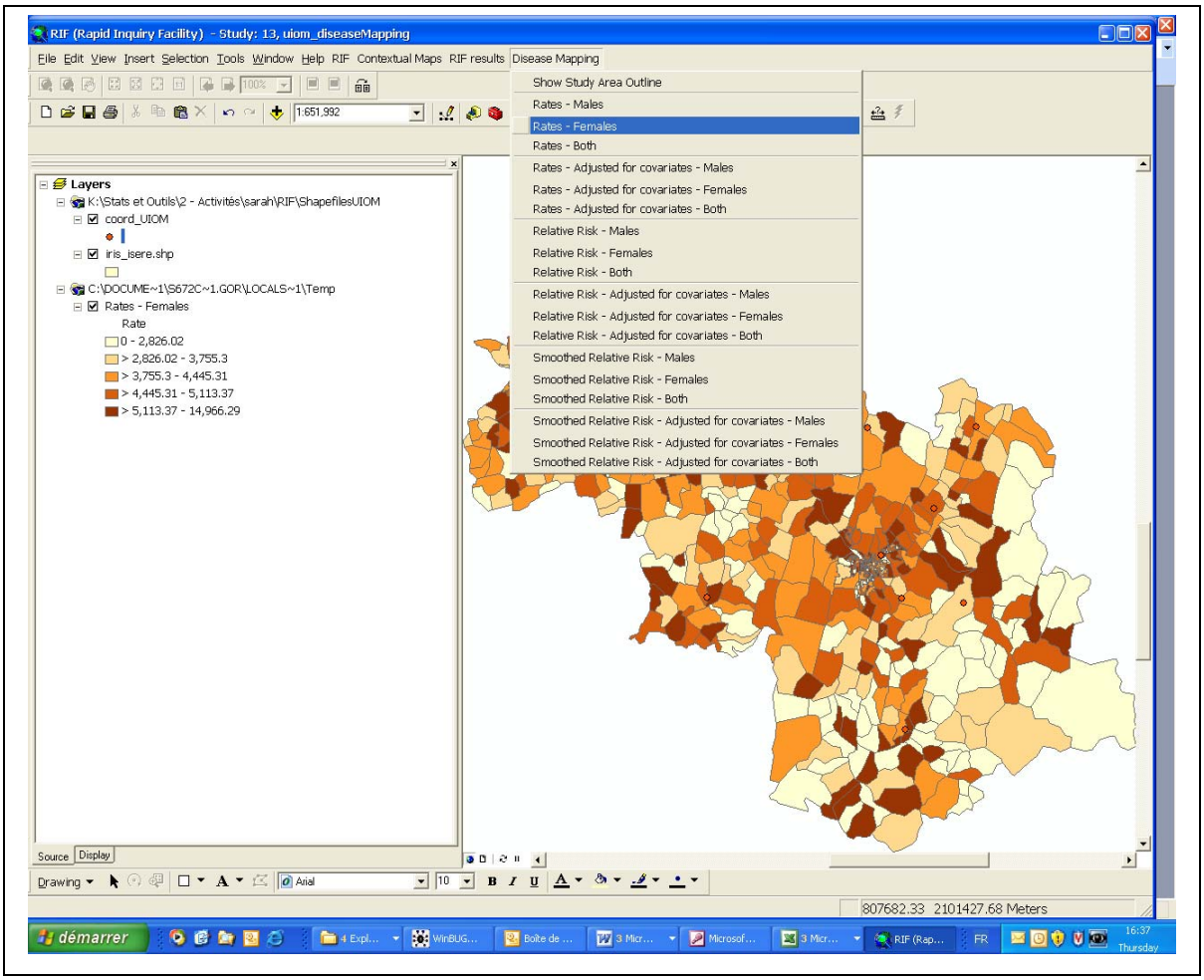
I FIGURE 30 I

Exemple de représentation des résultats de l'analyse autour de points sources dans RIF



I FIGURE 31 I

Exemple de représentations cartographiques disponibles dans RIF



4.2 Exemples d'utilisation de RIF

Les articles de Ball *et al.* [86] et Hodgson *et al.* [87] sont particulièrement intéressants quant à l'utilisation de RIF.

L'objectif de l'étude de Hodgson *et al.* est de déterminer s'il y a un risque plus important de développer une pathologie rénale chez les personnes vivant à proximité des industries de Runcorn, région North-West du Royaume-Uni. Le site industriel de Runcorn a été en activité pendant plus d'un siècle. Ce site, constitué de 16 industries, est responsable de la diffusion dans l'environnement (dans l'air et l'eau) de tonnes de produits chimiques : plomb, mercure, arsenic et chrome.

Les données de mortalité et morbidité sont analysées avec RIF. La mortalité est étudiée en utilisant des données recueillies en routine et fournies par l'Office for National Statistics. La morbidité est étudiée en utilisant les données d'admissions hospitalières locales recueillies en routine également. L'unité géographique est le district : la plus petite unité spatiale du recensement national. La distance aux industries est utilisée comme un proxy de l'exposition.

Les ratios standardisés de mortalité ajustés sur l'âge, le sexe et la défaveur sociale (indice de Carstairs) sont calculés pour les affections rénales (nephritis, nephritic syndrome et nephrosis) observés entre 1981 et 1999. La population de référence est la population de la région North-West. Ces taux sont calculés pour les districts dont le centroïde est entre 0 et 2 km et 2,01-7,5 km des installations industrielles.

Les ratios standardisés d'admissions ajustés sur l'âge et le sexe sont calculés pour les maladies rénales bénignes (non-malignant renal disease) et pour les cancers du rein observés pendant la période 1990-1999. Ces taux sont calculés par district et la population de référence est celle des villes de Warrington et Halton.

Un excès significatif de morbidité et de mortalité par maladie rénale est observé chez les personnes vivant à proximité des industries – vivant dans des zones potentiellement exposées aux produits toxiques émis par ces installations.

Les auteurs soulignent que deux sources de données indépendantes ont été utilisées et des résultats cohérents ont été obtenus. Les excès observés pourraient être liés à l'exposition aux produits toxiques émis par le site de Runcorn.

Néanmoins, les analyses faites dans cette étude (analyse écologique et cartographie des maladies) sont rudimentaires pour établir un risque sanitaire associé à une source de pollution. Il s'agit d'un travail d'investigation préliminaire sur la santé de la population vivant à proximité des industries de Runcorn. Il est donc nécessaire d'approfondir le lien entre les excès de risque observés et la pollution chimique des zones étudiées (pour pouvoir établir une relation causale).

Hodgson *et al.* [88] ont repris l'étude de 2004 mais en estimant l'exposition au mercure par modélisation de la dispersion atmosphérique (ADMS-Urban® version 2.0). Le modèle est validé par des mesures de concentration de mercure sur le site de Runcorn. Les SMR, ajustés sur l'âge (classes d'âge de cinq ans) et l'indicateur de Carstairs, sont calculés pour trois classes de la variable d'exposition au mercure.

Ball *et al.* [86] comparent différentes méthodes d'analyse de données de cancer. RIF et SaTScan® sont utilisés pour étudier le risque de cancer d'une population vivant à proximité d'une base aérienne à l'origine de la contamination au trichloréthylène d'une nappe phréatique dans l'Utah. Les localisations de cancers étudiées sont : le poumon, le rein et les lymphomes malins non-hodgkiniens.

Les données de cancer proviennent du registre de cancer de l'Utah. Six périodes consécutives de cinq ans de 1975 à 2004 sont étudiées. L'unité géographique est le census block : la plus petite unité du recensement national américain. Les contours des panaches contaminés au trichloréthylène et aux composés associés sont disponibles. Si un census block est à moins de 400 m d'un panache et d'un niveau de concentration de 5-10 µg/L, alors il est considéré potentiellement exposé. La zone d'étude non exposée donne la population de référence.

L'analyse est faite par âge et sexe et avec deux covariables supplémentaires, le niveau socio-économique (revenu médian du census block) et le pourcentage de la population qui est résidente depuis plus de cinq ans (une mesure de la mobilité de la population, un proxy de la durée potentielle d'exposition).

RIF est utilisé pour calculer les ratios standardisés pour la population potentiellement exposée pour chaque localisation de cancer et pour chaque période d'étude. La population potentiellement exposée est comparée à la population de référence.

RIF est utilisé aussi pour la représentation cartographique des cancers. Les SMR lissés et non lissés sont calculés. Le taux de référence est ici celui de la zone d'étude toute entière. L'outil de représentation cartographique des maladies facilite les analyses exploratoires et les clusters sont identifiés visuellement.

SaTScan® est aussi utilisé. Une analyse spatio-temporelle est faite pour détecter des agrégats circulaires ou elliptiques. L'incidence du cancer du poumon et l'incidence du cancer du rein sont significativement plus élevées chez la population potentiellement exposée pour deux des six périodes étudiées. Ces ratios sont calculés en prenant en compte les deux covariables supplémentaires.
La représentation cartographique du cancer du poumon obtenue avec RIF et les résultats de SaTScan® sont présentés. Deux clusters significatifs sont identifiés. Quand on inclut les deux covariables, aucun cluster n'est identifié avec SaTScan®.

Cette étude démontre l'utilité de RIF comme outil d'analyse rapide de risque de maladie dans une population définie et comme outil d'exploration de la distribution géographique d'une maladie en connexion avec SaTScan®.

Utiliser des méthodes qui permettent d'explorer la structure spatiale et temporelle d'une maladie aide à identifier des populations et des facteurs potentiellement d'intérêt pour des investigations ultérieures.

Dans cette étude, des excès de risque pour les cancers du poumon et du rein sont mis en évidence chez la population potentiellement exposée. Mais, cette analyse ne fait pas (et ne peut pas faire) le lien entre risque de cancer et exposition. L'inclusion de covariables qui pourraient expliquer ces excès de risque est indispensable.

RIF permet d'améliorer la capacité et l'efficacité d'investigations de santé publique de maladies liées à l'environnement comme le cancer.

L'étude de Ferrandiz *et al.* [89] est un autre exemple intéressant de l'utilisation de RIF. Ici, il est utilisé pour explorer l'association entre la mortalité de maladies cardiovasculaires et la "dureté" de l'eau de boisson.

D'une manière générale, il s'agit d'études descriptives qui ont pour objectif de décrire l'état de santé d'une population vivant à proximité d'une source polluante à partir de données disponibles, il s'agit en particulier de comparer

l'incidence/mortalité avec celle d'une population de référence et mettre en évidence ou pas une éventuelle surincidence/mortalité en relation avec un risque environnemental. L'idée est de s'appuyer sur des données recueillies en routine.

4.3 Développement de RIF

Une version mise à jour du RIF est prévue pour 2010. La mise à jour des données devrait être facilitée pour les études menées en routine. Des modèles statistiques pour l'analyse des éventuelles relations santé-environnement et des modèles statistiques pour la représentation spatio-temporelle des maladies devraient être mis en œuvre.

Le développement de RIF se réalise dans le cadre des projets EUROHEIS et EUROHEIS2 et du National Environmental Public Health Tracking Network du CDC.

Le National Institute for Public Health and the Environment (RIVM) est en train de développer un projet dans lequel l'outil RIF est un élément essentiel.

- **Les projets européens EUROHEIS (2000-2003) et EUROHEIS2 (2007-2010)**

L'objectif du projet EUROHEIS (http://www.euroheis.org/) était d'améliorer l'analyse de données sanitaires pour l'estimation des associations entre pollution environnementale et maladie et pouvoir répondre rapidement à des menaces environnementales en améliorant les connaissances et la compréhension de la gestion des risques sanitaires [90].

L'outil RIF était développé et mis en œuvre pour explorer les liens entre exposition à des polluants environnementaux et risques sanitaires potentiels. Il a été nécessaire de faire l'inventaire des bases de données géographiques existantes dans chaque pays participant au projet, de vérifier leur disponibilité, accessibilité et coût, et de recueillir des informations sur leur qualité et exhaustivité, ceci pour des données de santé, démographiques, environnementales et socio-économiques.

L'utilité de RIF était démontrée par des cas d'étude réalisés dans chaque pays participant au projet.

L'objectif du projet EUROHEIS2 est d'améliorer l'analyse, la communication et la diffusion d'information sur les risques sanitaires associés à des pollutions environnementales locales ou régionales.

Ce projet poursuit le développement du système d'information santé et environnement RIF débuté dans EUROHEIS. Le projet se focalise sur les outils et méthodes pour l'évaluation rapide des risques sanitaires liés à l'environnement. Un des enjeux est l'inclusion dans RIF des méthodes de représentation cartographique spatio-temporelle des maladies potentiellement liées à une exposition environnementale. Une des difficultés est la diversité des données des différents pays participant au projet.

- **Le national EPHT Network du CDC**

La technologie des SIG et les méthodes d'analyse spatiale associées sont au centre du système de surveillance de l'état de santé des populations aux États-Unis. L'équipe du SAHSU, en partenariat avec le CDC, adapte l'outil RIF à des programmes de surveillance sanitaire de certains états dans le cadre du programme national de surveillance en santé publique et environnement. Le but étant d'évaluer les relations spatio-temporelles entre une pathologie et une exposition environnementale.

- **Le programme Small Area Health Analyses (SMARHAGT) du RIVM**

L'objectif du programme SMARHAGT du RIVM est de développer un outil permettant la surveillance de la santé environnementale, l'analyse de groupement de cas, l'évaluation de risque à l'échelle nationale, régionale et locale à partir de données d'exposition et de santé disponibles en routine.

Les objectifs spécifiques sont :
- faciliter la représentation cartographique des maladies en utilisant RIF, pour explorer les liens entre exposition environnementale et indicateurs sanitaires ;
- faciliter des études de corrélation géographique ;
- faciliter l'utilisation des méthodes de détection de clusters spatiaux ou spatio-temporels ;
- construire des bases de données nationales avec des expositions environnementales, des indicateurs démographiques et socio-économiques géoréférencés pour être utilisés à une échelle locale.

4.4 Conclusion : utilité et limites de RIF

RIF a été développé pour répondre rapidement à des questions "environnement-santé". C'est un outil de description des données de santé. Il ne permet pas d'estimer une relation entre des pathologies et des sources de pollution mais de formuler des hypothèses, quant à un ou des facteurs explicatifs. Le lien direct avec des logiciels comme SaTScan® et WinBUGS® permet d'approfondir l'investigation et représente un atout supplémentaire.

Le développement d'un outil comme RIF a permis de prendre conscience des "possibles" de l'approche spatiale dans une étude épidémiologique. Aujourd'hui, d'autres programmes de santé publique développent des applications similaires adaptées à leurs besoins : par exemple, le programme national de surveillance en santé publique et environnement du CDC et le projet SMARHAGT du RIVM.

L'inconvénient de RIF est que la base de données est lourde à mettre en place (contraintes imposées par le logiciel). D'autre part, le tutoriel est relativement peu détaillé. Le fait que RIF dépende de l'acquisition du logiciel ARCGIS® d'ESRI© représente aussi une contrainte non négligeable même si aujourd'hui, le monde de la santé est quasi exclusivement équipé par ESRI©.

La mise en place d'un projet RIF doit faire l'objet d'un plan d'étude comme pour toute étude épidémiologique. S'agissant d'études locales, il faut avoir défini : la zone d'étude et la zone potentiellement exposée, la population de référence, l'unité géographique d'analyse.

En résumé, s'équiper de RIF peut s'avérer intéressant pour mener rapidement des études descriptives en l'absence de spécialistes SIG et de statisticiens. Lors d'investigations en santé environnement, il est rare que l'on s'arrête à des analyses descriptives. En général, des analyses statistiques mettant en œuvre des méthodes plus poussées sont nécessaires.

Par ailleurs, RIF s'appuie sur la mise en place d'une base de données reliée à un SIG et dont l'administration est lourde. Par conséquent, nous pensons que RIF doit être envisagé dans les études ayant une base de données sanitaires ne nécessitant pas de mises à jour régulières. Pour autant, l'objectif est d'exploiter cette base de données pour répondre à différentes investigations.

5. Conclusion

Ce document a été construit dans l'objectif de montrer les différents types d'études spatiales en santé environnement, de balayer un certain nombre de méthodes statistiques et SIG, de fournir des références bibliographiques diverses sur l'utilisation des statistiques spatiales et des SIG en santé. Au terme de ce travail, il semble intéressant de revenir sur ces différentes méthodes en proposant des axes de développement et d'amélioration.

L'approche spatiale en santé environnement permet d'améliorer la description d'un fait de santé en le replaçant dans son contexte environnemental, notamment lorsqu'un lien avec l'environnement est suspecté. Les méthodes d'analyse géographique mises en œuvre aussi bien au moyen des SIG que par les statistiques spatiales contribuent ainsi à développer la connaissance sur ces évènements sanitaires et sur les populations et les territoires qu'ils concernent. Ce faisant, elles s'inscrivent pleinement dans les missions qui sont celles de l'InVS. Mais les méthodes statistiques et les utilisations des SIG qui permettent de prendre en compte la répartition spatiale d'un fait de santé et d'un contexte environnemental sont nombreuses et leur mise en œuvre nécessite une réflexion approfondie.

Nous avons centré ce travail sur les études écologiques dans lesquelles l'on traite des données agrégées et non individuelles. Malgré un certain nombre de biais et de difficultés d'interprétation liés précisément à la nature agrégée des données, ces études présentent certains avantages, notamment en termes de puissance statistique, d'étendue de la zone et de la population d'étude. Elles peuvent aider à générer des hypothèses quant à l'effet d'une exposition sur la santé au niveau agrégé. Ces hypothèses permettent, dans un deuxième temps, d'amorcer une réflexion sur la causalité de cet effet. L'objectif de ces études est d'estimer les risques liés à la survenue d'événements rares, soit pour obtenir une représentation cartographique des risques, la plus informative possible, soit pour quantifier les liens entre un indicateur sanitaire et des covariables environnementales. De nombreux travaux sont consacrés au développement méthodologique des études écologiques géographiques en santé-environnement et concernent en particulier les méthodes de détection de clusters, les modèles spatiaux, spatio-temporels, les modèles conjoints de plusieurs maladies ou de sources de données multiples, la convergence des algorithmes MCMC et, plus généralement, l'estimation dans le contexte bayésien. Parallèlement, l'utilisation accrue des SIG en santé environnementale rend compte de son intérêt

dans les problématiques traitées. Il convient de poursuivre les développements méthodologiques avec l'objectif d'affiner les méthodes d'analyse mises en œuvre dans la construction des indicateurs d'exposition et des covariables. Il faut également encourager l'amélioration de la collecte des données en vue d'un géo-référencement de meilleure qualité.

Les projets européens tels que EUROHEIS et EUROHEIS2 et le développement d'outils tels que RIF ont permis de diffuser les études écologiques géographiques dans plusieurs Instituts de santé publique européens et au CDC, et de rendre les méthodes associées à ces études plus accessibles. Les programmes SMARHAGT du RIVM et le National EPHT du CDC ont fait le choix d'utiliser pleinement ces outils en les intégrant dans leurs réflexions en santé environnementale. Ces différents projets sont la preuve que, malgré les biais et les difficultés induits par les études géographiques, celles-ci trouvent aujourd'hui de plus en plus leur place dans le champ de l'épidémiologie environnementale.

Les travaux de développement méthodologique devront avoir comme objectif de tenter de réduire ces biais. L'apport des études de corrélations géographiques ne nous semble pas devoir être remis en cause, mais un certain nombre de conditions doivent être vérifiées : il faut notamment qu'une mesure de l'exposition soit disponible, qu'il soit possible de prendre en compte les facteurs de confusion potentiels, que l'unité spatiale soit adaptée aux objectifs de l'étude. Combiner des données d'exposition individuelles ou intra-unité spatiale avec les données agrégées permet d'améliorer ce type d'étude [17,18]. Par ailleurs, en général, dans les études de corrélations géographiques, les données de santé sont agrégées sur des périodes relativement longues (de 10 ans ou plus) et l'information temporelle n'est pas exploitée. Utiliser des modèles spatio-temporels permettrait d'ajouter une composante temporelle et d'observer d'éventuelles interactions spatio-temporelles.

Enfin, un des axes de ces développements concerne plus particulièrement les études d'investigations autour d'un point source, pour lesquelles il semble plus pertinent de réaliser une étude multicentrique autour de sites présentant les mêmes caractéristiques d'émission [52-54;91;92].

Nous avons décrit l'approche spatiale telle qu'elle est intégrée aujourd'hui dans les études que nous sommes amenés à conduire dans le cadre de notre travail au DSE de l'InVS. Nous continuerons à suivre les développements méthodologiques qui tendent à réduire les biais constatés dans les études écologiques.

De manière plus générale, la connaissance géographique, parce qu'elle implique la connaissance des territoires et des populations, semble trouver pleinement sa place dans les missions qui sont celles d'un institut comme l'InVS. Une telle réflexion géographique doit donc être de plus en plus souvent intégrée dans ses travaux, comme c'est le cas pour de nombreuses études de santé menées dans d'autres pays, et notamment dans les pays anglo-saxons.

6. Références bibliographiques

[1] Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. Environ Health Perspect 2008;116(8):1105-10.

[2] Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial epidemiology: methods and applications. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications.Oxford: Oxford University Press; 2000. p. 3-14.

[3] Goria S, Le Tertre A. Les études locales autour d'un point source - Les différentes méthodes statistiques, leurs avantages et leurs inconvénients. Note méthodologique. Saint-Maurice: Institut de veille sanitaire; 2010. 8 p. Disponible à partir de l'URL : http://www.invs.sante.fr

[4] IRSN. Les études épidémiologiques des leucémies autour des installations nucléaires chez l'enfant et le jeune adulte: revue critique. 2008.

[5] Lawson AB, Biggeri A, Williams FLR. A review of modelling approaches in health risk assessment around putative sources. In: Lawson AB, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R, (dir.). Disease mapping and risk assessment for public health. Chichester: Wiley; 1999. p. 231-45.

[6] Guihenneuc-Jouyaux C. Statistical modelization of geographic variations: a major challenge in epidemiology and statistics. Rev Epidemiol Santé Publique 2002;50(5):409-12.

[7] Richardson S. Problèmes méthodologiques dans les études écologiques santé-environnement. CR Acad Sci Paris, Sciences de la Vie/Life Sciences 2000;323:611-6.

[8] Best NG, Cockings S, Bennett JE, Wakefield JC, Elliott P. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. Journal of the Royal Statistical Society, Series A 2001;164:155-74.

[9] Cordier S, Chevrier C, Robert-Gnansia E, Lorente C, Brula P, Hours M. Risk of congenital anomalies in the vicinity of municipal solid waste incinerators. Occup Environ Med 2004;61(1):8-15.

[10] Maheswaran R, Haining RP, Pearson T, Law J, Brindley P, Best NG. Outdoor NOx and stroke mortality: adjusting for small area level smoking prevalence using a Bayesian approach. Statistical methods in medical research 2006;15(5):499-516.

[11] Nieuwenhuijsen MJ, Toledano MB, Bennett J, Best N, Hambly P, de HC *et al*. Chlorination disinfection by-products and risk of congenital anomalies in England and Wales. Environ Health Perspect 2008;116(2):216-22.

[12] Richardson S, Monfort C, Green M, Draper G, Muirhead C. Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. Stat Med 1995;14(21-22):2487-501.

[13] Fabre P, Daniau C, Goria S, de Crouy-Chanel P, Empereur-Bissonnet P. Étude d'incidence des cancers à proximité des usines d'incinération d'ordures ménagères. Saint-Maurice: Institut de veille sanitaire; 2008. 139 p. Disponible à partir de l'URL : http://www.invs.sante.fr

[14] Wakefield J. Disease mapping and spatial regression with count data. Biostatistics 2007;8(2):158-83.

[15] Salway R. Statistical issues in the analysis of ecological studies, Ph.D. Thesis Imperial College School of Medicine, University of London; 2003.

[16] Wakefield JC, Salway R. A statistical framework for ecological and aggregate studies. Journal of the Royal Statistical Society, series A 2001;164:119-37.

[17] Salway R, Wakefield J. A hybrid model for reducing ecological bias. Biostatistics 2008;9(1):1-17.

[18] Jackson C, Best N, Richardson S. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. Journal of the Royal Statistical Society, Series A 2008;171(1):159-78.

[19] Best N, Ickstadt K, Wolpert R. Spatial Poisson regression for health and exposure data measured at disparate resolutions. Journal of the American Statistical Society 2000;95:1076-88.

[20] Fortunato L, Guihenneuc-Jouyaux C, Tirmarche M, Laurier D, Hémon D. Misspecification of within-area exposure distribution in ecological Poisson models. Environ Ecol Stat 2009;16:341-53.

[21] Fleuret S, Thouez JP. Géographie de la santé, un panorama. Paris : Economica; 2007.

[22] Nuckols JR, Ward MH, Jarup L. Using geographic information systems for exposure assessment in environmental epidemiology studies. Environ Health Perspect 2004;112(9):1007-15.

[23] Béguin M, Pumain D. La représentation des données géographiques: statistique et cartographie. Armand Colin éd.; 1994. 192 p.

[24] Bertin J. Sémiologie graphique: les diagrammes, les réseaux, les cartes. Paris : EHESS; 1999.

[25] Jenks GF, Caspall FC. Error on choroplethic maps: definition, measurement, reduction. Annals of the Association of American Geographers 1971;61(2):217-44.

[26] Colonna M. Habilitation à diriger des recherches Université Joseph Fourier, Grenoble; 2006.

[27] Pumain D, Saint-Julien T. L'analyse spatiale, localisation dans l'espace. Armand Colin éd. Paris: 2008. 166 p.

[28] Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis 1995;27(4):286-306.

[29] Vandentorren S. Exposition environnementale à l'amiante chez les personnes riveraines d'anciens sites industriels et affleurements naturels. Étude cas-témoins à partir des données du Programme national de surveillance du mésothéliome. Saint-Maurice: Institut de veille sanitaire; 2009. Disponible à partir de l'URL : http://www.invs.sante.fr.

[30] Counil E, Daniau C, Isnard H. Étude de santé publique autour d'une ancienne usine de broyage d'amiante : le Comptoir des minéraux et matières premières à Aulnay-sous-Bois (Seine-Saint-Denis) - Pollution environnementale entre 1938 et 1975 : impacts sanitaires et recommandations. Saint-Maurice: Institut de veille sanitaire; 2007. 254 p. Disponible à partir de l'URL : http://www.invs.sante.fr.

[31] De Crouy-Chanel P. Étude SIG de la corrélation entre exposition indirecte à l'amiante et asbestose. Geomatique Expert 2007;54:28-32

[32] Poulstrup A, Hansen HL. Use of GIS and exposure modeling as tools in a study of cancer incidence in a population exposed to airborne dioxin. Environ Health Perspect 2004;112(9):1032-6.

[33] Yu CL, Wang SF, Pan PC, Wu MT, Ho CK, Smith TJ *et al.* Residential exposure to petrochemicals and the risk of leukemia: using geographic information system tools to estimate individual-level residential exposure. Am J Epidemiol 2006;164(3):200-7.

[34] Hoek G, Beelen R, de Hoogh K, Vienneau D, Gulliver J, Fischer P *et al.* A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 2008;42:7561-78.

[35] Best N, Ickstadt K, Wolpert R, Briggs D. Combining models of health and exposure data: the SAVIAH study. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications.Oxford: Oxford University Press; 2000. p. 393-414.

[36] Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. Hoboken, New Jersey: Wiley; 2004.

[37] Elliott P, Wakefield JC, Best NG, Briggs DJ. Spatial epidemiology: methods and applications. Oxford: Oxford University Press; 2000.

[38] Disease mapping with a focus on evaluation. Stat Med 19; 2000.

[39] Huang L, Pickle LW, Das B. Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. Stat Med 2008;27(25):5111-42.

[40] Demattei C. Détection d'agrégats temporels et spatiaux, Ph.D. Thesis Université Montpellier 1 UFR de médecine, Montpellier; 2006.

[41] Wakefield JC, Kelsall JE, Morris SE. Clustering, cluster detection, and spatial variation in risk. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications. Oxford: Oxford University Press; 2000. p. 128-52.

[42] Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med 1995;14(8):799-810.

[43] Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. Stat Med 2006;25(22):3929-43.

[44] Kulldorff M. A spatial scan statistic. Commun Stat Theory Methods 1997;26(6):1481-96.

[45] Kulldorff M. SaTScan User Guide for version 7.0; 2006.

[46] Bivand RS, Pebesma EJ, Gomez-Rubio V. Applied spatial data analysis with R. Springer; 2008.

[47] Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 2005;4:11.

[48] Assuncao R, Costa M, Tavares A, Ferreira S. Fast detection of arbitrarily shaped disease clusters. Stat Med 2006;25(5):723-42.

[49] Morris SE, Wakefield JC. Assessment of disease risk in relation to a pre-specified source. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, (dir.). Spatial epidemiology: methods and applications.Oxford: Oxford University Press; 2000;153-84.

[50] Bithell JF, Stone RA. On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. J Epidemiol Community Health 1989;43(1):79-85.

[51] Stone RA. Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. Stat Med 1988;7(6):649-60.

[52] Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford J *et al.* Cancer incidence near municipal solid waste incinerators in Great Britain. Br J Cancer 1996;73(5):702-10

[53] Bithell JF, Dutton SJ, Draper GJ, Neary NM. Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. BMJ 1994;309(6953):501-5.

[54] White-Koning ML, Hemon D, Laurier D, Tirmarche M, Jougla E, Goubin A *et al.* Incidence of childhood leukaemia in the vicinity of nuclear sites in France, 1990-1998. Br J Cancer 2004;91(5):916-22.

[55] Gomez-Rubio V, Ferrandiz-Ferragud J, Lopez-Quilez A. Detecting clusters of disease with R. Journal of Geographical Systems 2005;7(2):189-206.

[56] Bithell JF. The choice of test for detecting raised disease risk near a point source. Stat Med 1995;14(21-22):2309-22.

[57] Kulldorff M. Tests of spatial randomness adjusted for an inhomogeneity: a general framework. Journal of the American Statistical Association 2006;101(475):1289-305.

[58] Tango T. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med 1995;14(21-22):2323-34.

[59] Tango T. A test for spatial disease clustering adjusted for multiple testing. Stat Med 2000;19(2):191-204.

[60] Chirpaz E, Colonna M, Viel JF. [Cluster analysis in geographical epidemiology: the use of several statistical methods and comparison of their results]. Rev Epidemiol Sante Publique 2004;52(2):139-49.

[61] Thomas A, Best N, Lunn DJ, Arnold R, Spiegelhalter D. GeoBUGS User Manual; 2004.

[62] Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 1987;43(3):671-81.

[63] Breslow NE, Clayton DG. Approximate inference in generalised linear mixed models. Journal of the American Statistical Association 1993;88:9-25.

[64] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 2000;10:325-37.

[65] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics (with Discussion). Annals of the Institute of Statistical Mathematics 1991;43(1):1-59.

[66] Bernardinelli L, Clayton DG, Pascutto C, Montomoli C, Ghislandi M, Songini M. Bayesian analysis of space-time variation in disease risk. Stat Med 1995;14(21-22):2433-43.

[67] Waller LA, Carlin BP, Xia H, Gelfand AE. Hierarchical spatio-temporal mapping of disease rates. Journal of the American Statistical Association, 1997;92:607-17.

[68] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. Stat Med 2000;19(17-18):2555-67.

[69] MacNab YC, Dean CB. Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. Biometrics 2001;57(3):949-56.

[70] MacNab YC, Dean CB. Spatio-temporal modelling of rates for the construction of disease maps. Stat Med 2002;21(3):347-58.

[71] Abellan JJ, Richardson S, Best N. Use of space-time models to investigate the stability of patterns of disease. Environ Health Perspect 2008;116(8):1111-9.

[72] Ugarte MD, Goicoa T, Ibanez B, Militino AF. Evaluating the performance of spatio-temporal Bayesian models in disease mapping. Environmetrics 2009;20:647-65.

[73] Rue H, Martino S, Chopin N. Approximate bayesian inference for latent gaussian models by using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, series B 2009;71:319-92.

[74] Green PJ, Richardson S. Hidden Markov models and disease mapping. Journal of the American Statistical Association 2002;97:1055-70.

[75] Richardson S, Thomson A, Best NG, Elliott P. Interpreting posterior relative risk estimates in disease mapping studies. Environ Health Perspect 2004;112:1016-25.

[76] Best N, Hansell AL. Geographic variations in risk: adjusting for unmeasured confounders through joint modelling of multiple diseases. Epidemiology 2009;20(3):400-10.

[77] Dabney AR, Wakefield JC. Issues in the mapping of two diseases. Statistical methods in medical research 2005;14:83-112.

[78] Held L, Natario I, Fenton SE, Rue H, Becker N. Towards joint disease mapping. Statistical methods in medical research 2005;14:61-82.

[79] Tzala E, Best N. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. Statistical methods in medical research 2008;17:97-118.

[80] Richardson S. Spatial models in epidemiological applications. In: Green PJ, Hjort NL, Richardson S, (dir.). Highly Structured Stochastic Systems.Oxford: Oxford Statistical Science Series; 2003. p. 237-59.

[81] Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 2005;14(1):35-59.

[82] Latouche A, Guihenneuc-Jouyaux C, Girard C, Hemon D. Robustness of the BYM model in absence of spatial variation in the residuals. Int J Health Geogr 2007;6:39.

[83] Lee DJ, Durban M. Smooth-CAR mixed models for spatial count data. Computational Statistics and Data Analysis 2009;53:2968-79.

[84] Thurston SW, Wand MP, Wiencke JK. Negative binomial additive models. Biometrics 2000;56(1):139-44.

[85] Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics (with discussion). Appl Statist 1998;47:299-350.

[86] Ball W, LeFevre S, Jarup L, Beale L. Comparison of different methods for spatial analysis of cancer data in Utah. Environ Health Perspect 2008;116(8):1120-4.

[87] Hodgson S, Nieuwenhuijsen MJ, Hansell A, Shepperd S, Flute T, Staples B *et al.* Excess risk of kidney disease in a population living near industrial plants. Occup Environ Med 2004;61(8):717-9.

[88] Hodgson S, Nieuwenhuijsen MJ, Elliott P, Jarup L. Kidney disease mortality and environmental exposure to mercury. Am J Epidemiol 2007;165(1):72-7.

[89] Ferrandiz J, Abellan JJ, Gomez-Rubio V, Lopez-Quilez A, Sanmartin P, Abellan C *et al.* Spatial analysis of the relationship between mortality from cardiovascular and cerebrovascular disease and drinking water hardness. Environ Health Perspect 2004;112(9):1037-44.

[90] Kokki E, Ranta J, Penttinen A, Pukkala E, Pekkanen J. Small area estimation of incidence of cancer around a known source of exposure with fine resolution data. Occup Environ Med 2001;58(5):315-20.

[91] Elliott P, Richardson S, Abellan JJ, Thomson A, de HC, Jarup L *et al.* Geographic density of landfill sites and risk of congenital anomalies in England. Occup Environ Med 2009;66(2):81-9.

[92] Evrard AS, Hemon D, Morin A, Laurier D, Tirmarche M, Backe JC *et al.* Childhood leukaemia incidence around French nuclear installations using geographic zoning based on gaseous discharge dose estimates. Br J Cancer 2006;94(9):1342-7.

Introduction aux statistiques spatiales et aux systèmes d'information géographique en santé environnement

Application aux études écologiques - Résultats 2010

Les variations spatiales des indicateurs de santé et des facteurs d'expositions environnementales sont étudiées en épidémiologie dans un but descriptif et afin de suggérer des hypothèses étiologiques.

L'objectif de ce travail est de présenter et discuter les principaux outils et méthodes mettant en œuvre des systèmes d'information géographiques (SIG) et les statistiques spatiales utilisées dans les études écologiques géographiques. Ce travail s'intéresse aux études écologiques dans lesquelles les variables (indicateurs de santé et facteurs de risque) sont mesurées à l'échelle d'une unité géographique (commune, îlots regroupés pour l'information statistique (Iris), etc.) et non à l'échelle de l'individu. Sont décrites et discutées des méthodes statistiques adaptées à l'analyse des relations entre indicateurs sanitaires et indicateurs d'exposition à des facteurs de risques environnementaux. La modélisation et l'analyse statistique de ces données posent un certain nombre de difficultés méthodologiques : la forte variabilité, la dépendance spatiale, l'existence de différentes échelles spatiales, etc. Sont présentés les outils statistiques les plus utilisés pour répondre à ces difficultés.

Les possibilités qu'offrent la mise en œuvre des SIG et l'exploitation des données géographiques sont présentées en s'appuyant sur des exemples concrets de travaux menés au Département santé environnement (DSE) de l'Institut de veille sanitaire (InVS) ainsi que quelques exemples issus de la littérature, en insistant sur les précautions qui doivent accompagner leur utilisation.

Mots clés : étude écologique géographique, système information géographique, statistiques spatiales, représentation cartographique des maladies, détection de clusters spatiaux

Introduction to spatial statistics and geographic information systems in environmental health

Application to ecological studies - Results 2010

Spatial variations of health indicators and factors of environmental exposures are studied in epidemiology for descriptive purposes and to suggest etiological hypotheses.

The objective of this study is to present and discuss the main tools and methods implementing geographic information systems (GIS) and the spatial statistics used in ecological and geographical studies. This work focuses on ecological studies in which variables (health indicators and risk factors) are measured at the scale of a geographical unit (county, census block, etc.) rather than on the individual level. Statistical methods adapted to analyzing relationships between health indicators and indicators of exposure to environmental risk factors are described and discussed. Modeling and statistical analysis of these data raise a number of methodological difficulties: high variability, spatial dependence, existence of different spatial scales, etc. The most widely used statistical tools to address these difficulties are presented.

The possibilities related to the GIS implementation and the operating of geographical data are presented based on concrete examples of activities conducted at the Department of Health and Environment of the French Institute for Public Health Surveillance, as well as some examples from the literature, emphasizing the precautions that must accompany their use.

Citation suggérée :
Goria S, Stempfelet M, de Crouy-Chanel P. Introduction aux méthodes statistiques et aux systèmes d'information géographique en santé environnement – Application aux études écologiques – Résultats 2010. Saint-Maurice: Institut de veille sanitaire; 2011. 65 p. Disponible à partir de l'URL : http://www.invs.sante.fr.

Institut de veille sanitaire
12 rue du Val d'Osne
94 415 Saint-Maurice Cedex France
Tél. : 33 (0)1 41 79 67 00
Fax : 33 (0)1 41 79 67 67
www.invs.sante.fr

ISSN : 1958 -9719
ISBN-NET : 9 78-2-11-128246-9
Dépôt légal : mars 2011