Journées de méthodologie statistique de l'Insee 2022

Modélisation statistique du lien entre l'exposition indirecte aux produits phytosanitaires agricoles et le risque de survenue d'une hémopathie maligne (HM) en France : proposition d'une méthode spatialisée

Caleb Carloss AGUIDA[1,2], *Hong-Phuong* DANG[3,4],
Alain MONNEREAU[1,2], *Blandine* VACQUIER[2], *Sebastien* ORAZIO[1,2]

[1] *Registre des Hémopathies Malignes de la Gironde, Institut Bergonié, 33000, Bordeaux, France*
[2] *Epicene team, University of Bordeaux, Iserm, Bordeaux Population Health Research Center, UMR 1219, 33000, Bordeaux, France*
[3] *Ensai, CREST - UMR 9194*
[4] *Univ. de Bretagne Occidentale, LaTIM, INSERM - UMR 1101*

s.orazio@bordeaux.unicancer.fr

Mots-clés. (6 maximum) : modèle bayésien, modèle spatialisé, pesticide, hémopathie maligne (HM), maladie rare.

Domaines. Autre, Santé-environnement

Résumé

Les derniers résultats des études épidémiologiques ont ravivé l'inquiétude des riverains des zones traitées par des pesticides d'autant plus que la France est l'un des premiers consommateurs Européen de pesticides agricoles. En France, il existe peu d'informations sur l'impact sanitaire des pesticides lié au voisinage d'activités agricoles recourant à l'épandage de pesticides en population générale (riverains). Le projet GEO-K-PHYTO propose à travers une étude écologique, d'évaluer l'existence d'un potentiel lien entre l'exposition indirecte aux pesticides et la survenue d'hémopathies malignes. Pour cela, des choix méthodologiques ont été faits pour définir le type d'indicateur sanitaire à utiliser, l'indicateur d'exposition indirecte aux pesticides adapté des données disponibles et la méthode d'analyse statistique la plus appropriée.

Nous proposons la mesure d'association classiquement utilisée en épidémiologie : le rapport standardisé d'incidence (ou Standardized Incidence Ratio (SIR)) comme indicateur sanitaire et la surface agricole utile (SAU) par habitant au m^2 habitable comme indicateur d'exposition indirecte. Afin d'estimer l'effet de l'exposition indirecte des riverains sur le SIR, nous proposons d'utiliser le modèle bayésien spatialisé BYM avec une vraisemblance de type Zero Inflated Poisson pour représenter l'excès de zéros afin de prendre en compte les nombreuses communes sans cas en raison de la rareté des maladies étudiées.

La mise en œuvre de cette proposition nous a permis d'obtenir des résultats qu'on peut toutefois améliorer en développant d'une part un nouvel indicateur plus précis en matière de proximité des riverains de parcelles agricoles, mais également en travaillant à une échelle géographique plus fine.

Abstract

The latest results of epidemiological studies have rekindled the concern of people living near areas treated with pesticides, especially since France is one of the leading European consumers of agricultural pesticides. In France, there is little information on the health impact of pesticides related to the vicinity of agricultural activities using pesticides in the general population (residents). The GEO-K-PHYTO project proposes, in a first ecological study, to evaluate the existence of a potential link between indirect exposure to pesticides and the occurrence of a hematological malignancy. For this, methodological choices were made to define the type of health indicator to be used, the indirect pesticide exposure indicator adapted to the available data and the most appropriate statistical analysis method.

We propose the measure of association classically used in epidemiology : the Standardized Incidence Ratio (SIR) as a health indicator and the useful agricultural area per capita per m^2 of living space as an indirect exposure indicator. In order to estimate the effect of indirect exposure of residents on the SIR, we propose to use the Bayesian spatialized model BYM with a Zero Inflated Poisson likelihood to represent the excess of zeros in order to take into account the numerous municipalities without cases due to the rarity of the diseases studied.

The implementation of this proposal has allowed us to obtain results that can be improved by developing a new and more precise indicator for the proximity of people living near agricultural plots, but also by working on a finer geographic scale.

1 Introduction

La France est le premier pays agricole Européen en termes de Surface Agricole Utile (SAU). Elle est également l'un des premiers consommateurs de pesticides agricoles en Europe [1]. Le modèle agricole Français, interdépendant de l'utilisation de produits chimique, a fait de notre pays le principal utilisateur de pesticides en Europe avec près de 58000 tonnes de produits « phytosanitaires » vendus en moyenne chaque année entre 2010-2020 (hors usage en agriculture bio et hors produits de biocontrôle) [2]. Depuis 2020, la consommation Française se situe autour de 44000 tonnes et même si cette consommation est dépassée par celle de l'Espagne, elle demeure l'une des plus importantes au monde. Sur la période 2009-2019, la tendance de l'exposition de la population Française aux produits chimiques agricoles au lieu de se réduire semble se renforcer. L'indicateur NODU (qui correspond au nombre de traitements appliqués à pleine dose sur une surface d'un hectare) est en constante augmentation [2].

Il est aujourd'hui établi que les pesticides épandus se propagent bien au-delà de la zone sur laquelle ils sont appliqués, du fait de phénomènes de dérive de pulvérisation et de volatilisation [3]. Les pesticides épandus se dispersent dans le sol, l'air et les eaux souterraines et de surface [4]. La présence de pesticides a été détectée dans toutes les phases atmosphériques, dans 91% des cours d'eau et dans 59% des eaux souterraines, à des taux parfois supérieurs à la concentration maximale autorisée [5]. Les pesticides les plus présents dans les eaux sont des herbicides [1]. Ainsi, les pesticides sont présents à proximité des populations riveraines des zones agricoles.

La dernière mise à jour de l'expertise collective de l'Inserm sur ce sujet [6] publiée en 2021 met en évidence des présomptions de liens forts entre ces pesticides et plusieurs maladies des travailleurs agricoles. Plus d'une vingtaine de pesticides ont été classifiés comme cancérogènes certains ou probables pour l'homme par le Centre International de Recherche sur le Cancers (CIRC) pour différentes localisations cancéreuses : testicule, prostate, foie et vésicule biliaire, sein, poumon et hémopathies malignes (HM). En ce qui concerne les HM (ensemble hétérogène de cancers des cellules sanguines et leurs précurseurs), depuis 2013, 49 nouvelles publications ont été analysées dans la dernière mise à jour de l'expertise collective de l'Inserm et renforce le niveau de preuves en particulier sur les Lymphomes non hodgkiniens ou certaines leucémies. [1].

A contrario, il existe peu d'études épidémiologiques sur l'impact sanitaire de l'épandage des pesticides du fait d'activités agricoles sur la population générale résidant à proximité de ces parcelles (effets sur les riverains). Le dernier plan cancer préconise donc de développer des travaux sur les effets de santé en lien avec l'exposition aux pesticides au regard du niveau de preuve scientifique actuel. Dans le cadre du projet GEO-K-PHYTO financé par le plan ECOPHYTO II+, nous souhaitons notamment étudier le lien entre la survenue d'une HM et une exposition indirecte des riverains résidant à proximité des pesticides agricoles.

En première phase, ce projet consiste en la réalisation d'une étude épidémiologique du type écologique à partir des informations déjà disponibles. Ces études n'apportent pas d'information sur l'estimation des risques au niveau individuel, mais recherchent d'éventuelles relations entre les variations spatiales de facteurs d'exposition environnementale et celles des indicateurs sanitaires. Il s'agit d'une étude de corrélation descriptive qui peut permettre de générer des hypothèses étiologiques individuelle [7]. Cependant, il est connu que ce type d'étude présente un biais écologique pouvant avoir des répercussions au niveau de l'interprétation des résultats [8].

Nous exposons ici notre approche méthodologique qui doit notamment nous permettre de répondre aux quatre enjeux suivants :

— Proposer un indicateur d'exposition indirecte des riverains proche de la réalité de l'exposition individuelle ;
— Prendre en compte l'hétérogénéité des taux d'incidence des hémopathies malignes (en

particulier la faible probabilité de survenue de l'événement par unité géographique) ;
— Tenir compte de la spatialisation des indicateurs (hétérogénéité/autocorrélation spatiale) ;
— Proposer une estimation des effets (visualisation de sa dynamique).

2 Étude écologique (ou corrélation géographique)

L'objectif des études écologiques est donc d'analyser, au niveau de groupes d'individus définis sur une base géographique, la relation entre un indicateur de santé et une exposition environnementale. Elles doivent être considérées comme des études descriptives pouvant conduire à des hypothèses étiologiques individuelles et statuer sur l'intérêt de poursuivre de futurs investigations plus fines mais plus coûteuses (de type cas-témoins par exemple).

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

2.1 Pathologies d'intérêts

Nous nous intéressons uniquement aux hémopathies malignes (HM) qui peuvent se définir simplement comme un ensemble hétérogène de cancers des cellules sanguines et de leurs précurseurs. La classification OMS 2016 des hémopathies malignes recense plus de 160 maladies différentes. Ces maladies ne sont pas encore systématiquement étudiées séparément en épidémiologie car elles demeurent relativement rares mais elles sont regroupées par grandes entités : la totalité des HM représente 12% des nouveaux cas de cancer en France sur l'année 2018. Cinq sous-types représentent néanmoins la moitié des HM le myélome multiple/plasmocytome (5 442 nouveaux cas), le lymphome diffus à grandes cellules B (LDGCB) (5 071), les syndromes myélodysplasiques (4 735), la leucémie lymphoïde chronique (LLC) / lymphome lymphocytique (4 674) et les leucémies aiguës myéloïdes (LAM) (3 428) [9].

Le tableau 1 donne les sous-types d'HM étudiés et fournit également les taux d'incidences estimés en France pour l'année 2018.

2.2 L'unité spatiale

En première approche nous avons fait le choix d'un découpage géographique de type administratif à la commune. Un découpage à l'Iris (découpage administratif des communes afin d'avoir au maximum 5000 personnes) a été envisagé, mais la disponibilité des données sanitaires avec cette précision à l'Iris n'était pas envisageable dans l'immédiat. La nature administrative du découpage risque d'amener une forte hétérogénéité dans la répartition démographique entre unités avec des zones peu peuplées (zones rurales) et des zones densément peuplées (zone urbaines).

2.3 La zone d'étude

Le choix de la zone d'étude est le résultat de la disponibilité des données sanitaires et des informations nécessaires au calcul de l'indicateur indirect d'exposition aux pesticides. Huit départements Français répondent à ces critères : la Charente-Maritime, les Deux- Sèvres, la Charente, la Vienne, la Loire-Atlantique, la Vendée et la Gironde (*cf.* figure 1). On comptabilise sur ces départements un total de 2546 communes qui constitueront nos unités spatiales. Cette zone d'étude présente des cultures diverses avec une forte prédominance de vigne pour la

TABLE 1 – Hémopathies malignes : taux d'incidence brut et standardisé monde (TSM), sexe-ratio, selon le sexe, en 2018, en France métropolitaine

	Codes morphologiques CIM-O3	Taux d'incidence brut[1]		Taux d'incidence standardisé [2]		Sexe ratio [3]
		H	F	H	F	H/F
LYMPHOME DE HODGKIN	9650/3 & 9655/3, 9659/3, 9661/3 & 9667/3	3,9	2,6	3,7	2,7	1,4
LYMPHOMES NON HODGKINIENS (LNH)						
LLC/Lymphome lymphocytique	9670/3, 9823/3	8,8	5,7	4	2,1	1,9
Lymphome folliculaire	(>= 9690/3 & <= 9698/3), 9597/3	5,3	4,2	2,9	2	1,5
Lymphome diffus à grandes cellules B	9678/3, 9679/3, 9680/3, 9684/3, 9688/3,9712/3, 9735/3, 9737/3, 9738/3	8,8	6,8	4,7	3,2	1,5
Lymphome à cellules du manteau	9673/3	2,1	0,6	1	0,2	5
Lymphome de Burkitt	9687/3, 9826/3	0,5	0,2	0,5	0,2	2,5
Lymphome de la zone marginale	9689/3, 9699/3	4,6	4	2,3	1,7	1,4
Myélome multiple / plasmocytome	(>= 9731/3 & <= 9734/3)	9	7,8	4,2	2,9	1,4
LLP/ M. de Waldenström	9761/3, 9671/3	2,8	1,3	1,2	0,5	2,4
Leucémie à tricholeucocytes	9940/3	0,8	0,2	0,5	0,1	5
Lymphome T/NK à cellules matures (LNH T)	(>= 9700/3 & <= 9719/3), 9827/3, 9831/3,9834/3, 9948/3, 9724/3, 9725/3, 9726/3	3,2	2,3	1,8	1,3	1,4
Leucémie/lymphome lymphoblastique à cellules précurseurs (B, T ou SAI)	9727/3, 9728/3, 9729/3, 9835/3, 9836/3,9837/3, (9811/3 &? 9818/3)	1,6	1,1	2	1,5	1,3
Leucémie aiguë myéloïde	9805/3, (>= 9806/3 & <= 9809/3),9840/3, (>= 9860/3 & <= 9874/3), (>= 9891/3 & <= 9931/3), 9984/3	5,7	4,9	3,1	2,3	1,3
HEMOPATHIES MYELOÏDES						
Leucémie myéloïde chronique (LMC)	9863/3, 9875/3	1,5	1,2	1	0,7	1,4
SMC autres que LMC	9950/3, 9960/3-9964/3	5,8	5,8	2,9	2,5	1,2
SYNDROMES MYÉLODYSPLASIQUES	9980/3, 9982/3, 9983/3, 9985/3,9986/3, 9989/3, 9991/3, 9992/3	9,2	5,5	3,4	1,6	2,1
LEUCÉMIE MYÉLOMONOCYTAIRE CHRONIQUE ET AUTRES SMM	9876/3, 9945/3, 9946/3, 9975/3	2,7	1,7	1,1	0,5	2,2

[1] : Nombre de nouveaux cas pour 100 000 personnes-années

[2] : Taux d'incidence standardisés sur la structure d'âge de la population mondiale et exprimés pour 100 000 personnes-années

[3] : Rapport homme / femme des taux standardisés d'incidence

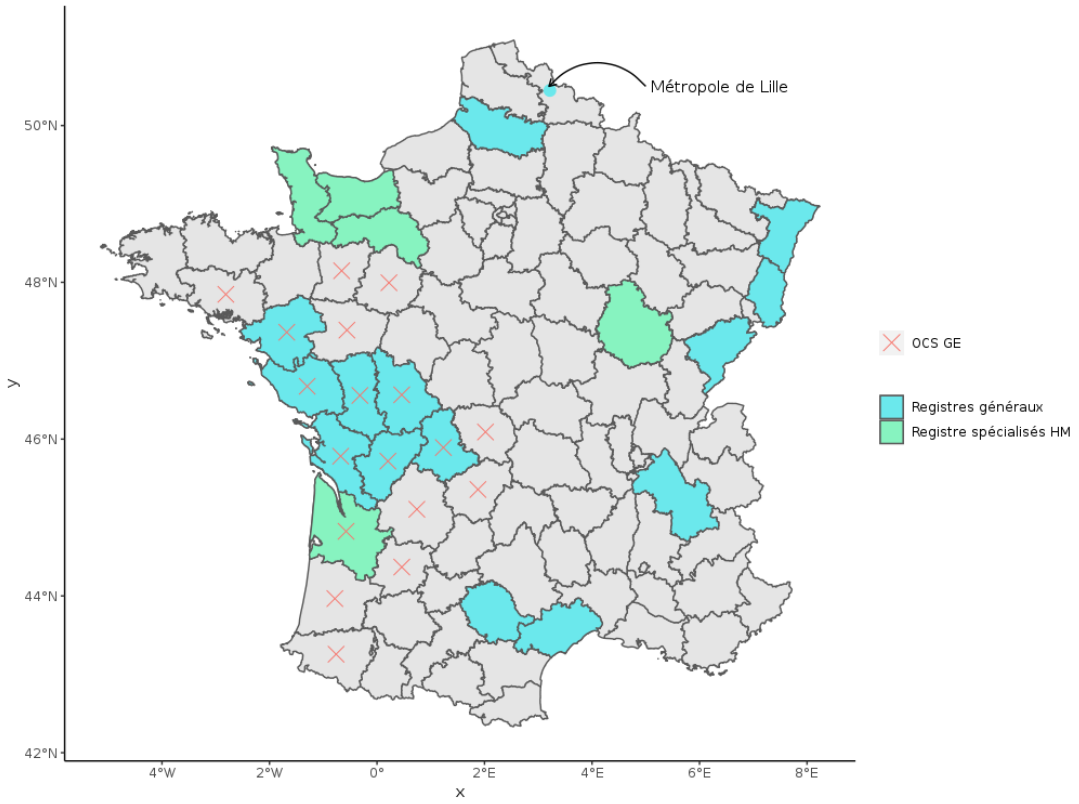


FIGURE 1 – Carte des Registres et de disponibilité des données d'occupation du sol à grande échelle (OCS GE disponible en 2019, avec deux années valides entre 2006 et 2017)

Gironde, et la présence de nombreuses « grandes cultures »(céréales) dans les départements de l'ex-région Poitou-Charentes.

2.4 Période d'étude et population d'étude

La période d'observation des cas d'HM a été définie de manière à faire porter l'étude sur un nombre important de personnes-années afin de disposer de suffisamment d'observation. La période d'étude est donc relativement longue et s'étend du 01/01/2006 au 31/12/2017.

L'étude porte sur des patients atteints de certains sous-type d'hémopathies malignes de l'adulte (plus de 15 ans révolus) diagnostiqués dans les départements de Charente-Maritime, Deux-Sèvres, Charente, Vienne, Loire-Atlantique, Vendée et la Gironde entre 01/01/2006 et le 31/12/2017.

2.5 Les facteurs de confusion

Richardson S. [10] nous rappelle que dans le cas d'études écologique et "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", la prise en compte de facteurs de confusion est un enjeu majeur. Au vu des informations disponibles nous avons pris en compte dans

nos analyses les facteurs de confusion suivant : la densité de population de la commune, le Zonage en Aires Urbaines (ZAU), l'indicateur écologique du niveau socio-économique de la commune (European Deprivation Indexe), le nombre d'industrie polluante et le potentiel Radon de la commune.

2.6 Les sources de données

Les données sanitaires de l'étude proviennent de la base commune FRANCIM qui regroupe les données du réseau français des registres des cancers. Les registres membres du réseau sont tous qualifiés par le Comité d'évaluation des Registres (CER) et garantissent l'exhaustivité et la qualité des informations recueillies par l'harmonisation des procédures d'enregistrement et de contrôle des données.
Une extraction des données de la base commune fournit pour chaque cas inclus les informations suivantes : le département, le sexe, la topographie, la morphologie, la date de diagnostic (cas incident), la classe d'âge quinquennale et la commune du domicile au moment du diagnostic.
Les populations à risque sont les habitants des communes constituants les huit départements de l'étude. Les données en accès de libre de l'INSEE mettent à disposition les estimations de population communale par classe d'âge et par sexe.
Les données nécessaires au calcul des indicateurs d'expositions indirectes aux produits phytopharmaceutiques sont fournies par l'IGN. L'institut a mis en place un système de calcul des métriques nécessaires aux analyses à partir d'un SIG (QGIS) et d'un ETL (FME). Ils exploitent un type particulier de base de données, celle de l'occupation du sol à grande échelle (OCS GE). A l'heure actuelle, seules les OCS GE produites par croisement de données et complétées par photo-interprétation sont éligibles. En France, il s'agit des seules informations permettant de définir, avec exhaustivité et finesse, le type de culture présente à l'échelle de la parcelle.

L'IGN met à disposition du projet, par commune, les informations suivantes : le département, la commune, la surface en m^2 par type de culture (vigne, verger, terre arable), la surface en m^2 de la commune ainsi que les polygones (shape).

Les informations sur les facteurs de confusions sont obtenues auprès de l'INSEE, de la plateforme MapInMed ou encore sur le site data.gouv.fr.

2.7 Indicateur indirect d'exposition

Le facteur de risque que nous souhaitons étudier est l'exposition indirecte des populations riveraines aux pesticides agricoles. Le calcul de la surface agricole utile (SAU) par commune est un proxy qui est couramment utilisé comme indicateur indirect de cette exposition dans les études écologique [1]. Cependant, il s'agit d'un indicateur peu représentatif de la probabilité individuelle d'exposition des populations ciblées. Le pourcentage de SAU est sensible aux effets de taille/densité de population de nos unités spatiales. Par exemple, on observe pour les communes avec une superficie agricole élevée, des populations à risque faible générant de l'instabilité dans les SIR. Dans d'autres situations, une commune avec 90% de surface agricole utile sera considérée comme étant fortement exposée alors qu'en réalité, on n'y retrouve que peu d'espace habitable le plus souvent concentrées dans des zones non exposées. Pour pallier ces phénomènes, on propose de calculer la SAU_{net} qui estimera le % SAU par habitant et par m^2 habitable. Il s'agit d'un lissage du % SAU pour tenir compte de l'espace potentiellement habité, de la taille de la commune et des effectifs de population à risque.

$$\%\text{SAU}_{net} = \%\text{SAU} \times \text{DENSITE}_{reelle} \quad \text{avec} \quad \text{DENSITE}_{reelle} = \frac{\text{Effectif de la commune}}{\text{Supercifie} \times (1 - \%SAU)}$$

Cet indicateur est plus proche de la probabilité individuelle d'exposition de nos unités spatiales que le % de SAU classiquement utilisé.

2.8 Définition de l'indicateur sanitaire : SIR

Le SIR (Standardized Incidence Ratio) est un indicateur souvent utilisé en épidémiologie pour quantifier un excès de l'incidence d'une maladie dans une population à partir d'un taux de référence. Il est défini par la formule suivante :

$$SIR = \frac{O}{E}$$

avec O (Observed) le nombre de cas observés de la maladie et E (Expected) le nombre de cas attendus. O est une quantité observée et E une quantité calculée dont le détail est donné dans le point suivant.

Calcul des attendus

La structure de la population étant différente d'une commune à une autre, les comparer à travers des indicateurs comme le taux d'incidence ou le SIR requiert préalablement de neutraliser l'effet structure de la population (la standardisation). La standardisation permet de corriger le déséquilibre entre les populations à comparer, en utilisant les taux spécifiques d'une population de référence.

Pour le choix de la population de référence, la population nationale (France) et la population de la zone d'étude(les huit (08) départements) ont fait l'objet de réflexion. Cependant, c'est le taux sur la zone d'étude qui sera utilisé car la zone d'étude n'est pas représentative de l'ensemble de la zone registre (qui a permis de faire l'estimation des taux nationaux) et donc utilisé un taux national n'est pas adapté. De plus, les taux nationaux pour certains sous types d'hémopathies malignes ne sont pas disponibles. Les SIR de la zone d'étude sont calculés avec une standardisation sur l'âge et le sexe. Il est connu que ces deux facteurs démographiques influencent le risque de survenue des hémopathies malignes. En standardisant sur l'âge et le sexe, nous gommons leurs effets étiologiques. En supposant $i \in S = \{1, 2\}$ l'indice représentant le sexe et $j \in \{1, \ldots, L\}$, l'indice de la tranche d'âge ([15 ;19] ;[20 ;24] ;... ;[85 ;++]), L étant le nombre de tranche d'âge, le taux spécifique de la zone d'étude par sexe et par tranche d'âge est donnée par :

$$t_{ij} = \frac{O_{ij}}{PA_{ij}}$$

où O_{ij} représente le nombre de cas observé de sexe i et de classe d'âge j ; PA_{ij} l'effectif personne-année de sexe i et de classe d'âge j de la population de la zone d'étude. Le calcul de l'attendu sur une commune (E^c) est défini alors par :

$$E^c = \sum_{i \in S} \sum_{j=1}^{L} N_{ij}^c \times t_{ij}$$

avec N_{ij}^c l'effectif de population de sexe i et de classe d'âge j dans la commune c.

Avec le SIR, les unités avec des petits effectifs peuvent avoir une variance associée aux SIR très grande, rendant les estimations du risque instable. La variabilité des SIR est différente selon les unités spatiales ce qui peut donner des SIR extrêmes correspondant le plus souvent aux unités les moins peuplées [11]. Cette instabilité est dû au fait de considérer les risques indépendamment, d'une unité à l'autre, sans prendre en compte la corrélation spatiale [7]. Des méthodes de lissage des SIR ont été développées pour produire des estimations plus fiables. L'intérêt du lissage est de permettre de mieux apprécier les structures spatiales sous-jacentes en lissant le bruit causé par l'instabilité des SIR dans les zones à petit nombre de cas.

3 Méthode d'analyse

3.1 Choix du modèle

Afin d'estimer l'effet de l'exposition indirecte des riverains sur le SIR, nous proposons d'utiliser le modèle bayésien spatialisé BYM [12] avec une vraisemblance du type Zero Inflated Poisson. Ce choix est guidé par la nature des données (excès de zéro) et aussi de la possibilité de prendre en compte l'autocorrélation spatiale. En effet, dans les études écologiques, l'autocorrélation spatiale est récurrente. L'inclusion de l'autocorrélation spatiale est utile car autrement les modèles classiques supposeraient une indépendance spatiale entre les unités statistiques généralement géographiques et cela pourrait entraîner des biais dans les résultats. Ignorer donc l'autocorrélation spatiale reviendrait à peu près à ignorer l'ordre des données dans une série temporelle.

3.2 Description du modèle BYM

Le modèle BYM cherche à modéliser le SIR des unités spatiales de l'étude. On peut montrer que modéliser une variable réponse sous une forme de ratio avec une distribution de poisson est équivalente à modéliser le numérateur avec une distribution de poisson en rajoutant le logarithme du dénominateur aux covariables.

$$\log\left(\frac{\mathbf{E}(O)}{E}\right) = \mu + X^\top\beta \Leftrightarrow \log(\mathbf{E}(O)) = \mu + X^\top\beta + \log(E) \tag{1}$$

avec μ l'intercept, β les coefficients des covariables et X ($\in \mathcal{X} \subseteq \mathbb{R}^p$) les variables explicatives. on rappelle que O (Observed) est le nombre de cas observés et E (Expected) le nombre de cas attendus. Des effets non linéaires des covariables peuvent être inclus dans le modèle par le biais des splines cubiques naturelles ou des fonctions de base cubiques des covariables dans X.

Le modèle BYM nous permet d'intégrer dans un modèle GLM (Generalized Linear Models) [13] la notion d'autocorrélation spatiale. Pour cela, on introduit dans l'équation (1) deux composantes aléatoires : ϕ une composante spatiale et θ une composante à effets aléatoires ordinaires pour l'hétérogénéité non spatiale. On peut réécrire l'équation (1) comme suit :

$$\log(\mathbf{E}(O)) - \log(E) = \mu + X^\top\beta + \phi + \theta = t + \phi + \theta. \tag{2}$$

La composante ϕ traduit l'idée que les valeurs pour une paire de zones contiguës seraient généralement beaucoup plus semblables que pour deux zones quelconques. On pose $x = t + \phi + \theta$.

En l'absence d'autres informations, on suppose que ϕ et θ sont indépendants et que θ est un bruit blanc gaussien de variance λ.[12]
Pour ϕ, on choisit une densité parmi la famille

$$p(\phi) \propto \exp\left\{-\sum_{i>j} w_{ij} f(\phi_i - \phi_j)\right\}, \phi \in \mathcal{R}^n \tag{3}$$

ici les $w_{ij} = \begin{cases} 0 & \text{si i et j sont des zones voisines} \\ 1 & \text{sinon.} \end{cases}$, et f(z) est une fonction paire et croissante en | z | .
La densité conditionnelle des ϕ_i est alors :

$$p(\phi_i|\phi_{-i}) \propto \exp\left\{-\sum_{j\in E_i} w_{ij} f(\phi_i - \phi_j)\right\}, \phi_i \in \mathcal{R} \tag{4}$$

où E_i est l'ensemble des voisins de la zone i et w_{ij}= w_{ji}.
Pour la fonction f, deux cas sont proposés :

$$f(z) = \begin{cases} z^2/2\kappa & (1) \\ |z|/\kappa & (2) \end{cases}$$

avec κ une constante inconnue strictement positive. En se plaçant dans le cas (1) avec $z = \phi_i - \phi_j$, on a :

$$p(\phi) \propto \exp\left\{-\frac{1}{2\kappa}\sum_{i\sim j}(\phi_i - \phi_j)^2\right\}, \phi_i \in \mathcal{R} \tag{5}$$

où $i \sim j$ implique que les zones i et j sont contiguës et $i > j$. Il s'agit d'un ICAR.

Généralement, les paramètres κ et λ ne sont pas connus et sont considérés aléatoires et indépendantes de β. Autrement dit, nous n'avons pas d'informations claires sur les paramètres κ et λ à part le signe de la valeur qu'ils peuvent prendre. On supposera dans ce document que κ et λ sont indépendantes et de densité : $\mathrm{p}(\kappa,\lambda) \propto \mathrm{e}^{-\epsilon/2\kappa}\mathrm{e}^{-\epsilon/2\lambda}$, κ, λ>0 et $\epsilon \approx 0.01$. Ainsi $\kappa \sim$ Gamma-Inverse$(0, 0.01)$ et $\lambda \sim$ Gamma-Inverse$(0, 0.01)$. La loi a priori de $\boldsymbol{\beta} = (\mu, \beta)$ est choisie gaussienne de moyenne 0 et de variance $\Sigma = \sigma^2 I_{p+1}$. Dans les calculs, on prendra $\sigma^2 = 10^5$. Nous désignons par la suite $Y(\in \mathcal{Y} \subseteq \mathbb{N})$ comme la variable de comptage du nombre de cas observés.
La densité a priori de $(\phi, \theta, \kappa, \lambda)$ est donnée par :

$$p(\phi,\theta,\kappa,\lambda) \propto \kappa^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\kappa}\sum_{i\sim j}(\phi_i - \phi_j)^2\right\} \times \lambda^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\lambda}\sum_{i=1}^{n}\theta_i^2\right\} \times \mathrm{p}(\kappa,\lambda) \tag{6}$$

Avec une vraisemblance de poisson, la densité postérieure conjointe de ϕ, θ, λ et κ, $\boldsymbol{\beta}$ est :

$$\begin{aligned} p(\phi,\theta,\kappa,\lambda,\boldsymbol{\beta}|\mathbf{Y},\mathbf{X}) \propto \prod_{i=1}^{n} \exp\left(\underbrace{-E_i \mathrm{e}^{\mu+\beta^\top X_i+\phi_i+\theta_i}}_{:=E_i h}\right)\left(E_i \mathrm{e}^{\mu+\beta^\top X_i+\phi_i+\theta_i}\right)^{Y_i} \\ \times \overbrace{\kappa^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\kappa}\sum_{i\sim j}(\phi_i - \phi_j)^2\right\}}^{:=A} \\ \times \overbrace{\lambda^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\lambda}\sum_{i=1}^{n}\theta_i^2\right\} \times \mathrm{p}(\kappa,\lambda) \times \mathrm{p}(\boldsymbol{\beta})}^{:=B}. \end{aligned}$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)$ et $\mathbf{X} = (X_1, \ldots, X_n)$.
Les densités conditionnelles :
- $\kappa|\theta,\phi,\lambda,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}$

$$p(\kappa|\theta,\phi,\lambda,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}) \propto \kappa^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\kappa}\sum_{i\sim j}(\phi_i - \phi_j)^2 - \epsilon/2\kappa\right\}$$

- $\lambda|\theta,\phi,\kappa,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}$

$$p(\lambda|\theta,\phi,\kappa,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}) \propto \lambda^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\lambda}\sum_{i=1}^{n}\theta^2 - \epsilon/2\lambda\right\}$$

Les densités conditionnelles de κ et λ appartiennent à la famille des distributions gamma-inverses.

- $\phi_i|\phi_{-i},\theta,\lambda,\kappa,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}$ pour $i=1,\ldots,n$

En remarquant que $\sum_{i\sim j}(\phi_i-\phi_j)^2 = \frac{1}{2}\sum_{i=1}^{n_i}\sum_{j\neq i}(\phi_i-\phi_j)^2 = n_i\sum_{i=1}^{n_i}(\phi_i-\overline{\phi}_i)$, avec $\overline{\phi}_i = n_i^{-1}\sum_{i=1}^{n}\phi_i$ on a :

$$p(\phi_i|\phi_{-i},\theta,\lambda,\kappa,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}) \propto \exp(-E_i e^{\mu+\beta^\top X_i+\phi_i+\theta_i})(E_i e^{\mu+\beta^\top X_i+\phi_i+\theta_i})^{y_i} \times \kappa^{-\frac{n}{2}} \exp\left\{-\frac{n_i}{2\kappa}(\phi_i-\overline{\phi}_i)^2\right\}$$
$$\propto \kappa^{-\frac{n}{2}} \exp\left(-E_i e^{\mu+\beta^\top X_i+\phi_i+\theta_i} + \phi_i y_i - \frac{n_i}{2\kappa}(\phi_i-\overline{\phi}_i)^2)\right)$$

- $\theta_i|\theta_{-i},\phi,\lambda,\kappa,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}$ pour $i=1,\ldots,n$

$$p(\theta_i|\theta_{-i},\phi,\lambda,\kappa,\boldsymbol{\beta},\mathbf{Y},\mathbf{X}) \propto \lambda^{-\frac{n}{2}} \exp\left(-E_i e^{\mu+\beta^\top X_i+\phi_i+\theta_i} + \theta_i y_i - \frac{1}{2\lambda}\theta_i^2\right)$$

- $\mu|\beta,\theta,\phi,\kappa,\lambda,\mathbf{Y},\mathbf{X}$

$$p(\mu|\beta,\theta,\phi,\kappa,\lambda,\mathbf{Y},\mathbf{X}) \propto \exp\left\{-\frac{1}{2\sigma^2}\mu^2 - \sum_{i=1}^{n} E_i e^{\mu+\beta^\top X_i+\phi_i+\theta_i} + \mu\sum_{i=1}^{n} Y_i\right\}$$

- $\beta_j|\beta_{-j},\mu,\theta,\phi,\kappa,\lambda,\mathbf{Y},\mathbf{X}$ pour $j=1,\ldots,p$

$$p(\beta_j|\beta_{-j},\mu,\theta,\phi,\kappa,\lambda,\mathbf{Y},\mathbf{X}) \propto \exp\left\{-\frac{1}{2\sigma^2}\beta_j^2 - \sum_{i=1}^{n} E_i e^{\mu+\beta^\top X_i+\phi_i+\theta_i} + \sum_{i=1}^{n} Y_i\beta_j X_{i,j}\right\}$$

Cas du ZIP : Prise en compte de l'inflation de zéros

Dans les données de comptage et surtout dans les études épidémiologiques de maladie rares, il y a souvent une grande proportion de zéros. On distingue deux types de zéros : ceux qui sont dûs à l'échantillonnage (zéros aléatoires) et ceux qui sont dûs à la structure (zéros structurels). Ne pas tenir compte de ce facteur peut conduire à un cas particulier de surdispersion, l'inflation de zéros (voir [14]). Ce phénomène a particulièrement été mis en évidence dans le cadre de la régression de Poisson et plusieurs outils ont été développé pour en tenir compte. L'une des approches les plus utilisées consiste à considérer un mélange de deux modèles. Cette approche nous conduit aux modèles à inflation zéros. De manière générale, un modèle à inflation de zéros est un mélange entre une distribution dégénérée en zéro et une distribution de comptage standard (par exemple Poisson, Binomial, Binomial négatif). Dans le cas du ZIP (Zero-inflated Poisson), la distribution de comptage est Poisson.
Supposons que $Y_i \sim \text{ZIP}(h,p)$, le modèle peut s'écrire :

$$\begin{cases} P(y_i=0) = p + (1-p)\exp(-E_i h) \\ P(y_i) = (1-p)\dfrac{\exp(-E_i h)(E_i h)^{y_i}}{!y_i} \qquad , y_i > 0 \end{cases}$$

où $0<p<1$. En désignant par $P(y_i;E_i h) = \dfrac{\exp(-E_i h)(E_i h)^{y_i}}{!y_i}$, alors

$$L' = \prod_{i=1}^{n} p\delta_0(y_i) + (1-p)P(y_i;E_i h).$$

On introduit une variable latente telle que : $\begin{cases} z_i = 0 & \text{si } y_i \text{ suit la dirac en zéro} \\ z_i = 1 & \text{sinon} \end{cases}$.

On a :

$$L_o = \prod_{i=1}^{n} [p\delta_0(y_i)]^{1-z_i} [(1-p)P(y_i; E_i h)]^{z_i}.$$

Dans le cadre d'une régression, p_i dépends des covariables $\mathbf{V}$ ($\in \mathcal{V} \subseteq \mathbb{R}^q$) avec q le nombre de variable explicatives pour l'inflation zéro. On pose alors

$$p_{v_i}(\omega_0, \omega) = \frac{1}{1 + \mathrm{e}^{-\omega_0 - V_i\omega}}$$

La postérieure s'écrit :

$$p(\phi, \theta, \kappa, \lambda, \boldsymbol{\beta}, \boldsymbol{\omega} | \mathbf{Y}, \mathbf{X}, \mathbf{V}) \propto \prod_{i=1}^{n} \left[\frac{1}{1 + \mathrm{e}^{-\omega_0 - V_i\omega}} \delta_0(y_i) \right]^{1-z_i} \left[(1 - \frac{1}{1 + \mathrm{e}^{-\omega_0 - V_i\omega}}) P(y_i; E_i h) \right]^{z_i} \times A \times B$$

avec $\mathbf{Y} = (Y_1, \ldots, Y_n)$; $\mathbf{X} = (X_1, \ldots, X_n)$ et $\mathbf{V} = (V_1, \ldots, V_q)$.
Pour obtenir la distribution postérieure, nous avons utilisés la méthodes Monte-Carlo par Chaines de Markov et plus précisément, l'échantillonneur de Gibbs. Cependant, les lois conditionnelles étant de formes complexes et non explicitement connues, elles ne sont pas directement simulables. Alors pour la mise à jour des paramètres à chaque itération de l'échantillonneur de Gibbs, l'algorithme de Metropolis-Hastings à marche aléatoire [15] est mis à contribution. Il est implémenté dans le package CARBayes [16] sous R (fonction S.CARbym()).

3.3 Chaînes MCMC et estimations

Nous avons effectué des simulations MCMC pour 1.000.000 d'itérations. Nous avons supprimé les 500 premières itérations considérées comme une période de rodage de la chaîne. Puis, dans les itérations restantes, nous avons conservé les tirages par pas de 2000 itérations afin de réduire la dépendance sérielle au sein des tirages. À la fin de ce post-traitement de notre échantillonneur MCMC, une chaîne de longueur d'environ 500 pouvant être raisonnablement considérée comme échantillonnée à partir de la distribution postérieure et présentant une très faible dépendance sérielle a été utilisée. Pour vérifier la convergence des chaînes MCMC obtenues, nous avons procédé par l'approche de diagnostic graphique(tracé des graphiques des chaînes pour chaque paramètre , tracé de la distribution conditionnelle correspondante, étude de la corrélation, etc...) .

3.4 Estimation flexible : Transformation non linéaire

Nous l'avons vu en introduction, nous souhaitons capter la dynamique des risques en fonction des niveaux d'exposition mesurés en continue. Il est impératif d'être en mesure de capter un potentiel effet non-linéaire de l'impact de notre indicateur d'exposition sur les RR.

Les splines permettent en général dans les modèles de régression de modéliser une relation non linéaire. Une fonction spline est un polynôme par morceaux à jonction lisse de degré n. Par jonction lisse, on entend que la fonction et ses $n-1$ premières dérivées doivent être continues aux points de jonction, ou nœuds. De manière générale, on peut écrire une fonction spline d'une variable x comme suit :

$$f(x) = \sum_{j=0}^{n} \beta_{oj} x^j + \sum_{i=1}^{K} \beta_{in} (x - s_i)_+^n \tag{7}$$

où s_i, $i = 1 \ldots n$ représente les noeuds et $h_+ = h$ si $h > 0$, sinon $h_+ = 0$.

On distingue $K + n + 1$ coefficients de régression pour cette fonction spline de degré n avec K nœuds. L'un des coefficients représente le terme constant.

Les splines cubiques sont fréquemment utilisées parce qu'elles offrent une grande flexibilité pour l'ajustement des données, qu'elles sont visuellement lisses, en raison de leurs dérivées premières et secondes continues, et qu'elles incluent moins de constantes pour l'ajustement que les splines de degré supérieur. Les splines naturelles sont des cas particuliers des splines cubiques. Elles sont contraintes à être linéaires dans les queues c'est à dire que $f(x)$ doit être linéaire pour $x < s_1$ et $x > s_k$. La première condition pour $x < s_1$ implique que $\beta_{02} = \beta_{03} = 0$. De même, la linéarité de $f(x)$ sur $x < s_k$ implique que $\sum_{i=0}^{k} \beta_{i3} = 0$ et $\sum_{i=1}^{k} \beta_{i3} s_i = 0$. Avec ces conditions, sur la base de l'équation (7) sans le terme constant, la transformation d'une variable x à l'aide d'une spline cubique restreinte avec k nœuds (c'est-à-dire $k - 1$ degrés de liberté, pour $k \geq 3$) aux emplacements $s_1, s_2, ..., s_k$ peut être écrit comme :

$$f(x) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1} \tag{8}$$

où $x_1 = x$ et pour tout $i = 1, \ldots, k-2$,

$$x_{i+1} = (x - s_i)_+^3 - \frac{s_k - s_i}{s_k - s_{k-1}}(x - s_{k-1})_+^3 + \frac{s_{k-1} - s_i}{s_k - s_{k-1}}(x - s_k)_+^3.$$

Par manque d'information précise sur les noeuds possibles dans notre étude, nous utilisons les noeuds par défaut de la fonction *ns()* du package *splines* sous R. Ces emplacements sont basés sur les percentiles de notre variable d'exposition.

4 Résultats et éléments de discussion

Finalement, la base de données de l'étude est constituée par 40659 cas d'hémopathies malignes de 15 ans et plus diagnostiqués entre le 01/01/2006 et le 31/12/2017 et répartis sur les 2546 communes constituant notre zone d'étude. Seuls seront présentés dans la suite du document les résultats nécessaires à la compréhension des choix méthodologiques qui ont été réalisés. La présentation des résultats finaux est soumise à l'approbation du conseil scientifique du projet GEO-K-PHYTO et des agences d'État impliquées.

4.1 Maladie rare et hétérogénéité des unités spatiales

La rareté des HM se traduit par une inflation du nombre de commune sans enregistrement de cas (*cf.* Table 2). Les sous-types présentant les meilleures situations comme par exemple le Myélome multiple / plasmocytome et la leucémie lymphoïde chronique / Lymphome lymphocytique comptabilise un peu moins de 45% de communes sans cas.

Les données d'études présentent un déséquilibre en ce qui concerne la répartition du nombre de cas par sous types d'hémopathies malignes (exemple du Myélome multiple, figure 2). L'unité géographique utilisée dans ce projet étant la commune, l'existence de grands pôles urbain (aux densités de population les plus fortes) entraîne une asymétrie dans la distribution du nombre de cas. Par exemple, la majorité des communes ont un nombre de Myélome Multiple / Plasmocytome situé entre 0 et 10, alors que quelques grands pôles urbains enregistrent un nombre de cas extrême (plus de 100 cas pour trois d'entre elles). A l'opposé, les communes aux SAU les plus fortes (+ de 25% d'exposition vigne dans notre exemple) comptabilisent pas ou très peu de cas et ont des densités de population parmi les plus faibles (exemple du %SAU vigne ici).

4.2 Indicateur d'exposition

La figure 3 illustre les différences en termes d'intensité d'exposition de l'indicateur proposé versus le % SAU classique. Nous observons des différences importantes dans le classement

TABLE 2 – Nombre de cas et nombre de communes recensant au moins un cas par sous-types d'hémopathies malignes diagnostiqués sur la période 2006 – 2017 et sur la zone d'étude (8 départements Français 33, 87, 79, 86, 17, 16, 86 et 44)

Hémopathies malignes	**Sous types**	**Nombre de Cas (n=42385)**	**Nombre de Communes***
Lymphome de Hodgkin	Lymphome de Hodgkin	1945	791
Lymphome non hodgkinien	Leucémie lymphoïde chronique / Lymphome lymphocytique	5363	1384
	Myélome multiple et plasmocytome	5092	1341
	Lymphome diffus à grandes cellules B	4837	1278
	Lymphome folliculaire	2807	947
	Lymphome de la zone marginale	2319	904
	Lymphome lymphoplasmocytaire / Macroglobulinémie de Waldenström	1962	795
	Lymphome T/NK à cellules matures (LNH-T)	1792	755
	Lymphome à cellules du manteau	770	434
	Leucémie / Lymphome lymphoblastique à cellules précurseurs (B, T ou SAI)	520	327
	Leucémie à tricholeucocytes	313	227
	Lymphome / Leucémie de Burkitt	170	130
Hemopathies myeloïdes	Syndromes myélodysplasiques	4589	1266
	Syndromes myéloprolifératifs chroniques (SMC), autres que LMC (autres SMC)	3244	1015
	Leucémies aiguës myéloïdes (LAM)	2957	992
	Leucémie myélomonocytaire chronique et autres SMM	1147	572
	Leucémie myéloïde chronique (LMC)	832	451

* : Nombre de Communes avec au moins 1 cas (n=2546)

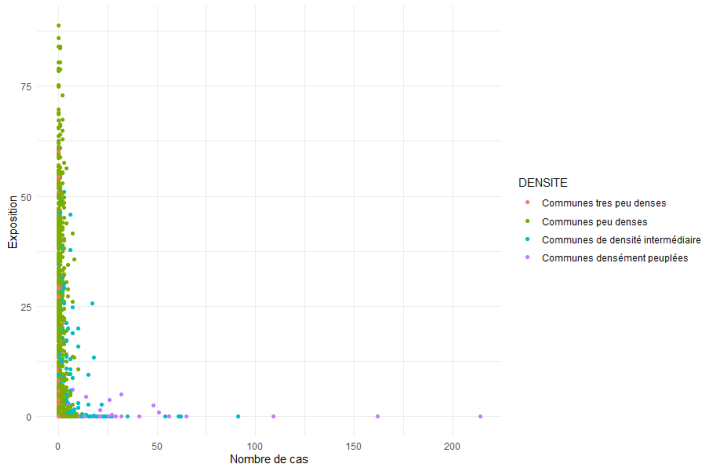
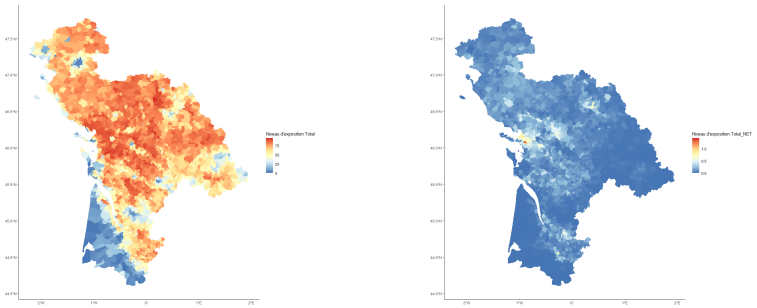


FIGURE 2 – Distribution du nombre de cas observés de Myélome Multiple / Plasmocytome en fonction de la surface agricole de vigne (noté « Exposition ») sur la période 2006 – 2017 et sur la zone d'étude (8 départements Français 33, 87, 79, 86, 17, 16, 86 et 44).



(a) Surface agricole utile brut (b) Surface agricole utile nette

FIGURE 3 – Cartographie des indicateurs d'expositions (Pourcentage de SAU brute vs pourcentage de SAU nette) résumé sur la période 2006-2017 et calculé à l'échelle de la commune.

des communes vis-à-vis de l'indicateur utilisé. Certaines unités spatiales à faibles densités de population (zone rurale), mais à fort %SAU se voient attribuées des expositions nettes très faible. C'est particulièrement visible pour les communes de la Haute-Vienne. A contrario, des zones urbaines avec des densités de populations concentrées autour de parcelles de vignes (Exemple : commune de Pessac) ont des expositions nettes non négligeables.

4.3 Modèles spatiaux et interprétation des effets flexibles

Les modèles BYM sont ajustés sur l'indicateur d'expositions indirectes des riverains (SAU nette) ainsi que sur les facteurs de confusion potentiels pour lesquels les informations sont disponible (densité de la commune, le Zonage en Aires Urbaines (ZAU), l'EDI, le nombre d'industrie polluante, le radon). Trois Surfaces Agricole Utile sont explorées séparément : vignes, vergers et terres arables. Les résultats indiquent que la convergence vers la distribution stationnaire, c'est-à-dire la distribution postérieure, peut-être supposée sans risque. Un extrait du diagnostic de la convergence est fourni en annexe 5. La variable d'exposition est modélisée comme non linéaire en utilisant une spline naturelle avec 2 degrés de liberté. Dans la pratique, pour l'interprétation des résultats obtenus à l'aide des splines, l'interprétation des coefficients estimés pour chaque paramètre utilisé dans les splines est essentiellement dénuée de sens. Pour cela, on a utilisé des combinaisons linéaires de ces coefficients pour obtenir les valeurs prédites et interpréter les résultats en utilisant ces prédictions. La meilleure façon d'interpréter les résultats des splines est d'utiliser une représentation graphique [17]. La figure 4 montre la dynamique du risque de survenue d'une hémopathie maligne en fonction de l'évolution du niveau d'exposition (la référence étant l'exposition=0). Sur chaque graphique, les estimations et les intervalles de crédibilité ponctuels à 95% sont indiqués.
Pour une interprétation ponctuelle, l'impact des co-facteurs étant égales par ailleurs ,

$$\begin{cases} \log\left(\text{Nombre de cas observé}_{ajt_0}\right) = \mu + \text{exposition}_0 + \ldots & \text{(a)} \\ \log\left(\text{Nombre de cas observé}_{ajt_k}\right) = \mu + (\text{exposition}_0 + \delta_k) + \ldots & \text{(b)} \end{cases}$$

donc

$$(b) - (a) \Rightarrow \log\left(\frac{\text{Nombre de cas observé}_{ajt_k}}{\text{Nombre de cas observé}_{ajt_0}}\right) = \delta_k \tag{9}$$

L'axe des y est le risque relatif (RR) estimé et l'axe des x le niveau d'exposition. Le risque relatif a l'interprétation suivante : les communes qui ont été exposées ont $(RR-1)\times 100\%$ de risque en plus (ou en moins) de survenue d'une HM par rapport aux communes qui n'ont pas été exposées (référence exposition=0).

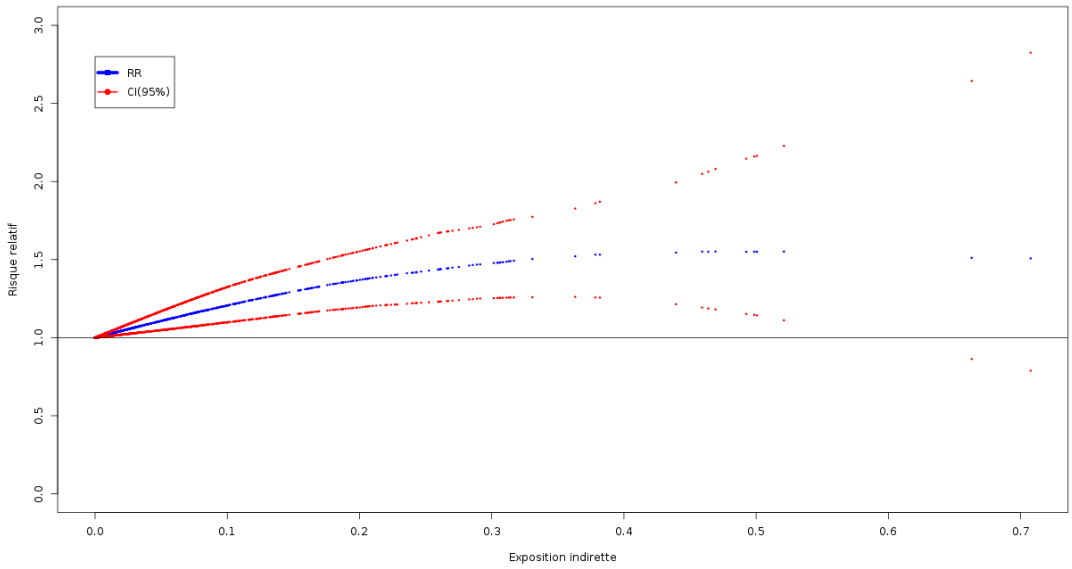


FIGURE 4 – Exemple de dynamique du risque relatif avec son intervalle de crédibilité

5 Intérêt et limites de l'approche

La méthodologie exposée ici, comme toute étude écologique, présente plusieurs avantages : elle est simple, rapide à mettre en œuvre et elle est peu coûteuse en raison de la disponibilité des données exploitées. Cependant plusieurs difficultés dans la mise en œuvre et l'interprétation de ces études sont à prendre en considération. D'une manière générale, ce travail n'offre qu'une fenêtre d'observation et de ce fait ne pourra pas servir à établir avec certitude un quelconque lien causal.

5.1 Distribution du nombre de cas

Les données d'études présentent un déséquilibre en ce qui concerne la répartition du nombre de cas par sous types d'hémopathies malignes. L'unité géographique utilisée dans ce projet étant la commune, l'existence de grand pôle urbain entraîne une hétérogénéité dans la distribution du nombre de cas. Enlever ces communes reviendrait non seulement à travailler sur un sous-ensemble de notre cohorte mais aussi sur une forte concentration de commune sans cas (lié au fait que l'on travaille avec des maladies rares). Par conséquent, les grands pôles urbains vont se comporter comme des valeurs extrêmes et à l'inverse, pour les petites et moyennes communes, on est confronté à un nombre élevé de 0 (pas de cas). Une solution serait d'utiliser dès que disponible les données par Iris.

Pour la structure spatiale, nous avons été contraint de supprimer les communes formant des îles (ex : l'île de ré) car elles n'ont pas de frontière commune avec les autres unités d'agrégation. Cependant la perte d'information n'a été que minime (20 communes ont été concernées sur un total de 2546).

5.2 L'indicateur d'exposition

Comme toute étude de ce type, nous sommes soumis à des biais écologiques qui est la différence potentiel entre le lien dose-effet individuel et celui estimé au niveau du groupe. Le biais écologique est dû à la variabilité intra-unité de l'exposition et des facteurs de confusion. En proposant un nouvel indicateur de type %SAU par habitant au m^2, nous cherchons à réduire cette

variabilité intra-unité. Cependant, cet indicateur d'exposition proposé bien qu'utile fait l'hypothèse forte que chaque individu est exposé de la même façon sur tout le territoire communal quelque soit sa localisation. Or, nous savons qu'en pratique ce n'est pas le cas. Par exemple, une maison en centre bourg ne sera pas du tout exposé car il n'y a aura aucune parcelle agricole à proximité. Il serait intéressant, par la suite, de réfléchir à un nouvel indicateur capable d'intégrer les données de géolocalisation des populations dans les unités géographiques afin de prendre en considération leur proximité réelle par rapport aux cultures.
Il serait également intéressant de renseigner pour chaque parcelle agricole le type et la quantité des produits phytosanitaires utilisés.

Également, dans ce type d'étude il est très difficile de prendre en compte des informations individuelles telle que l'historique résidentielle des patients. Or en l'absence d'informations sur l'immigration de la population, on suppose que la commune où le cas a été diagnostiqué est la commune où le patient a été exposé.

5.3 Prises en compte des facteurs de confusion

L'ajout dans le modèle de régression les facteurs de confusion adéquates est une approche pour maîtriser le biais écologique. En ce qui concerne les industries polluantes nous aurions pu aller plus loin qu'un simple comptage d'entreprise sur le territoire concerné. D'autres éléments pourraient être utilisés, comme la nature des polluants et/ou l'activité de l'entreprise. Certains polluants rejetés par les usines sont connus comme facteurs de risque des hémopathies malignes (ex : le benzène).

Également, il serait intéressant de prendre en compte d'autres facteurs individuels dont les effets cancérogènes sont bien établis comme les consommations de tabac et d'alcool. Une réflexion est en cours afin d'utiliser les données de ventes de ces deux biens de consommation dans les communes.

5.4 Choix des modèles

Une des difficultés relevées est intrinsèque à une approche utilisant des SIR dans le cas d'une maladie rare [18]. On observe fréquemment des unités spatiales sans cas et un risque d'instabilité des SIR (excès de risque apparent). Le modèle proposé avec le ZIP permet de corriger en partie ces difficultés.

En outre, l'utilisation du SIR pose un problème de biais de standardisation. En effet, la tranche d'âge et le sexe peuvent aussi impacter le niveau d'exposition. Donc,il peut être intéressant d'intégrer ces facteurs dans le modèle. On sait par ailleurs que certaines hémopathies par exemple n'interviennent qu'à l'âge adulte et d'autres beaucoup plus chez un sexe particulier.
Une alternative serait d'utiliser dans le cadre du BYM une vraisemblance du type Binomial au lieu du ZIP car le nombre maximal de cas de chaque unité géographique est conditionné à l'effectif de la commune. Cela nous permet de réduire l'espace des variables donc le temps de convergence. Cette approche pourrait nous permettre de s'affranchir du calcul des attendus et aussi de considérer directement l'âge et le sexe comme des covariables du modèle (voir annexe 6).

Une autre approche possible serait d'implémenter un modèle Binomial-Bêta avec une composante spatiale du type ICAR. On considère dans ce cas que la probabilité de succès (survenue d'une HM) n'est plus une fonction déterministe des covariables (contrairement à la précédente).

6 Conclusion

En conclusion, la méthodologie proposée semble raisonnable lorsque l'on cherche à estimer le risque de survenue d'une maladie relativement rare (par rapport au nombre de personne-année) en fonction de l'exposition indirecte des riverains aux pesticides. Cette première approche écologique doit permettre d'explorer rapidement les données et d'élaborer des hypothèses étiologiques individuelles.

Références

[1] Camille Roingeard, Alain Monnereau, Stéphanie Goujon, Sébastien Orazio, Ghislaine Bouvier, and Blandine Vacquier. Passive environmental residential exposure to agricultural pesticides and hematological malignancies in the general population : a systematic review. Environmental Science and Pollution Research, 28(32) :43190–43216, June 2021.

[2] Publication des données provisoires des ventes de produits phytopharmaceutiques en 2020 [Internet]. [cité 26 janvier 2022]. Disponible sur :. `https://agriculture.gouv.fr/publication-des-donnees-provisoires-des-ventes-de-produits-phytopharmaceutiques-en-2020.`

[3] Aaron Blair, Beate Ritz, Catharina Wesseling, and Laura Freeman. Pesticides and human health. Occupational and environmental medicine, 72, 12 2014.

[4] Jean-Noel J.-N. Aubertot, Jean Marc J. M. Barbier, Alain Carpentier, Jean-Noël Gril, Laurence L. Guichard, Philippe P. Lucas, Serge Savary, Marc Voltz, and Isabelle I. Savini. Pesticides, agriculture, environnement. Réduire l'utilisation des pesticides et en limiter les impacts environnementaux. Rapport. Other, INRA, 2005.

[5] INSERM. Pesticides. effets sur la santé. collection expertise collective,inserm, paris. Inserm, 2013.

[6] Inserm. Pesticides et effets sur la santé : Nouvelles donnée. Collection Expertise collective. Montrouge : EDP Sciences, 2021.

[7] C Guihenneuc-Jouyaux. Statistical modelization of geographic variations : A major challenge for epidemiology and statistical analysis. Revue d'épidémiologie et de santé publique, 50 :409–12, 11 2002.

[8] IRSN. Les études épidémiologiques des leucémies autour des installations nucléaires chez l'enfant et le jeune adulte : revue critique, 2.

[9] Defossez G, Le Guyader-Peyrou S, Grosclaude P Uhry Z, Colonna, Dantony E, and al. Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018. Volume 1 – Tumeurs solides. Saint-Maurice (Fra) : Santé publique France, 2019. 372 p.

[10] Sylvia Richardson. Problèmes méthodologiques dans les études écologiques santé environnement methodological problems in ecological studies of health environment effects. Comptes Rendus De L'Academie Des Sciences Serie Iii-sciences De La Vie-life Sciences, 2000.

[11] Colonna M. Habilitation à diriger des recherches université joseph fourier, Grenoble ; 2006.

[12] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics, 43(1) :1–20, 1991.

[13] P. McCullagh and J. A. Nelder. Generalized Linear Models. Chapman and Hall / CRC, London, 1989.

[14] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, 34(1) :1–14, 1992.

[15] Larissa Valmy. Modèles hiérarchiques et processus ponctuels spatio-temporels - Applications en épidémiologie et en sismologie. Theses, Université des Antilles-Guyane, November 2012.

[16] Duncan Lee. CARBayes : An R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software, 55(13) :1–24, 2013.

[17] Bryan E Shepherd and Peter F Rebeiro. Brief report : Assessing and interpreting the association between continuous covariates and outcomes in observational studies of hiv using splines. Journal of acquired immune deficiency syndromes (1999), 74(3) :e60 – e63, 2017.

[18] Bouyer J, Hémon D, Cordier S, Derriennic F, Stücker I, Stengel B, and al. Épidémiologie : principes et méthodes quantitatives, Paris : Inserm . 1995.

Annexe

Diagnostic de la convergence

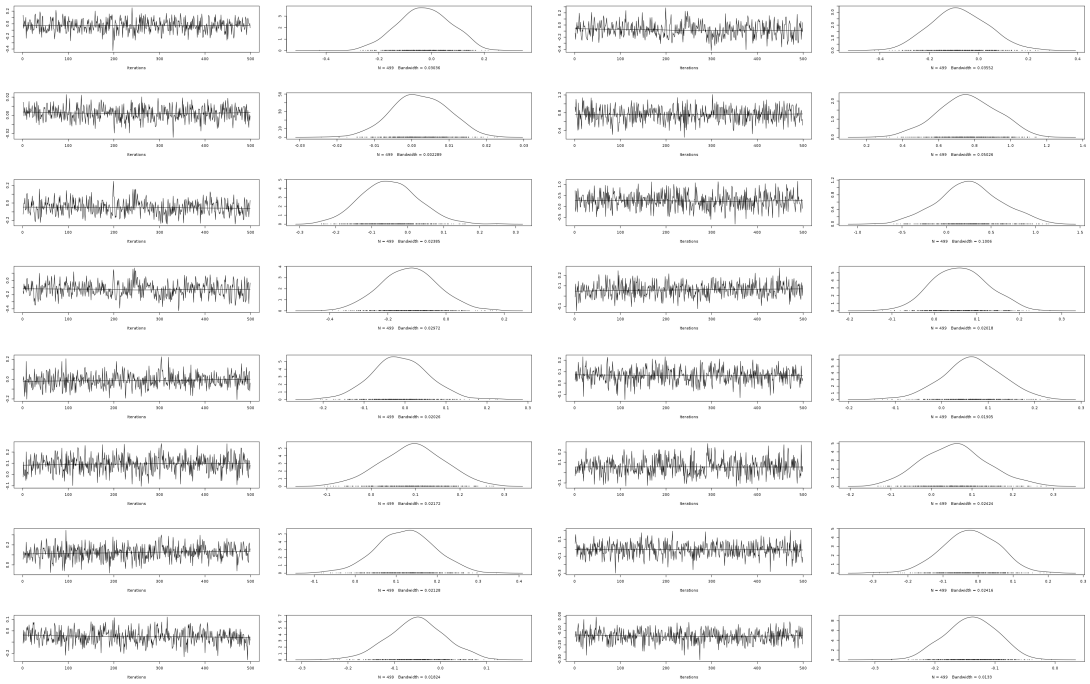


FIGURE 5 – Traceplot de la chaîne des variables explicatives

L'allure globale des historiques (voir 5, nombre d'itération = 1.000.000) ne montre aucune stagnation des chaînes autour d'une valeur pour tous les paramètres du modèle. Aucun motif ou périodicité n'apparaît dans la suite des valeurs. Toutes les valeurs semblent être régulièrement visitées.

En fixant une période de rodage de 500 (pour plus de sécurité) et en faisant un sous échantillonage à pas de 2000 observations (pour éliminer l'autocorrélation), nous avons vérifié que la longueur de la chaîne est assez grande pour assurer la stabilité des quantiles empiriques extrêmes des lois a posteriori (qui permettent de calculer les intervalles de crédibilité). Pour tous les paramètres du modèle, nous avons bien la stabilisation des quantiles ergodiques (voir 6).

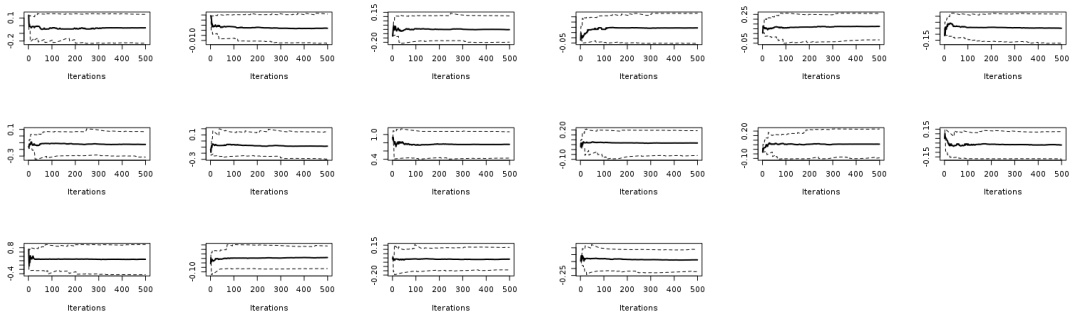


FIGURE 6 – cumuplot de la chaîne des variables explicatives

La taille de l'échantillon MCMC semble suffisante. On remarque également une décroissance assez rapide sur le graphique des auto-corrélations (voir 7).

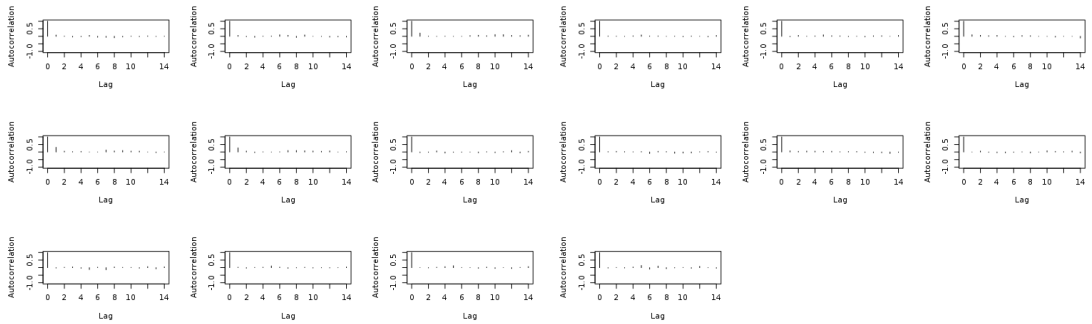


FIGURE 7 – Autocorrélation de la chaîne des variables explicatives

Le diagnostic de la convergence des chaînes montre bien qu'elles atteignent une convergence satisfaisante.

Le modèle statistique : 2e approche

On suppose Y_i le nombre de cas observé dans la commune i et m_i l'effectif de la commune i. On fait l'hypothèse que $Y_i \overset{\text{ind}}{\sim} \text{Binomial}(\pi_i, m_i)$.
Puisque $\pi_i \in (0,1)$, on peut utiliser la fonction de régression logistique pour introduire les covariables dans le modèle. On pose alors, $\pi_i = \dfrac{1}{1+\exp^{-\mu-\beta^\top X_i-\phi_i-\theta_i}}$
Avec une vraisemblance binomiale, la densité postérieure conjointe de ϕ, θ, λ et κ, $\boldsymbol{\beta}$ est :

$$p(\phi,\theta,\kappa,\lambda,\boldsymbol{\beta}|\mathbf{Y},\mathbf{X},\mathbf{m}) \propto \prod_{i=1}^{n}\left[\frac{1}{1+\exp^{-\mu-\beta^\top X_i-\phi_i-\theta_i}}\right]^{Y_i}\left[1-\frac{1}{1+\exp^{-\mu-\beta^\top X_i-\phi_i-\theta_i}}\right]^{m_i-Y_i} \times A \times B.$$

avec $\mathbf{Y} = (Y_1,\ldots,Y_n)$ et $\mathbf{X} = (X_1,\ldots,X_n)$ et $\mathbf{m}$ un vecteur de taille n contenant les effectifs des communes.