UNIVERSITÉ DE LILLE

FACULTE DE MÉDECINE HENRI WAREMBOURG

Année : 2020

THÈSE POUR LE DIPLÔME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Méthodes de détection de clusters spatiaux dans le cadre de la surveillance épidémiologique

Présentée et soutenue publiquement le 9 juin 2020 à 16h

Au Pôle Formation

Par Louis ROUSSELET

_____

JURY

Président :

Monsieur le Professeur Philippe AMOUYEL

Assesseurs :

Monsieur le Professeur Alain DUHAMEL

Monsieur le Professeur Emmanuel CHAZARD

Directeur de thèse :

Monsieur le Docteur Michaël GENIN

Travail du Laboratoire METRICS (ULR 2694) : Evaluation des technologies de santé et des pratiques médicales

_____

Avertissement

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

Liens d’intérêt

L’auteur et son directeur de thèse ne déclarent aucun lien d’intérêt en rapport avec le sujet traité au cours des trois années précédant la présentation de cette thèse.

Table des matières

1 Introduction générale

La surveillance épidémiologique est une discipline récente, apparue au milieu du XXe siècle, qui est devenue un outil important pour éclairer la prise de décision en santé publique (1). Elle permet de mesurer l'état de santé des populations et son évolution dans le temps et dans l'espace, mais aussi d'identifier les déterminants de la santé, qu'ils soient environnementaux, socio-économiques ou comportementaux (2). Elle repose parfois sur l'étude d'alertes lancées par des acteurs de terrain, professionnels de santé ou autre. Cela a été le cas à plusieurs reprises ces dernières années en France, notamment pour des agrégats de pathologies concentrés dans une commune ou un département (3). Ces alertes nécessitent confirmation, statistique d'abord car elles peuvent être dues au hasard, puis par une enquête plus poussée afin de vérifier les cas, puis d'en identifier la cause (4). Mais à chaque fois se pose la question de la capacité du système de surveillance épidémiologique à identifier ce type d'agrégats de façon proactive et non uniquement sur alerte venant du terrain (3). A ce jour, aucun système de détection d'agrégat n'est utilisé en routine dans ce cadre en France.

Un agrégat spatio-temporel (ou cluster) est défini comme « un regroupement dans le temps et dans l'espace de cas de maladies, de symptômes ou d'événements de santé au sein d'une population localisée » (5). Plusieurs types de tests permettent d'en détecter : tout d'abord l'approche globale qui permet de repérer une tendance globale à l'agrégation dans les données mais sans localiser les zones en cause dans cette tendance, ensuite les tests focalisés qui permettent d'évaluer la tendance à l'agrégation autour de points de l'espace fixés a priori, et enfin les méthodes de localisation de cluster qui permettent de les détecter sans a priori quant leur position (5). Ces méthodes ont été largement utilisées dans un cadre rétrospectif, mais plusieurs ont aussi été proposées pour un usage prospectif, afin de suivre régulièrement l'apparition de nouveaux agrégats.

Dans le cadre de cette thèse on se propose d'étudier la faisabilité de la mise en oeuvre d'un système de surveillance épidémiologique permettant de localiser des territoires nécessitant une intervention de santé publique. De tels territoires seront identifiés par la présence d'agrégat d'événements de santé dans l'espace et/ou dans le temps à l'aide de méthodes de détection de cluster prospectives.

Ce travail se présente en quatre parties :

La première s'attache à présenter le concept de surveillance épidémiologique ainsi que les grandes notions qui lui sont associées. Les différents objectifs possibles d'un système de surveillance sont évoqués, de même que les différents types de système pouvant être mis en oeuvre. Le processus de surveillance pouvant mener à une alerte sanitaire et les étapes de sa validation sont présentés. Puis, quelques exemples de systèmes de surveillance épidémiologique existants en France et à l'étranger sont exposés. Enfin, l'actualité récente de ce domaine nous conduit à identifier un besoin en matière de détection d'agrégats d'événements de santé et à définir les caractéristiques principales d'un système desurveillance pouvant répondre à ce besoin.

La deuxième partie se concentre sur la présentation des méthodes statistiques de la sur-

veillance épidémiologique. Dans un premier temps, sont présentées quelques méthodes communément utilisées dans le cadre de la surveillance de séries chronologiques utiles pour la suite. Sont ensuite évoquées les méthodes de surveillance permettant la détection de clusters spatio-temporels d'événements. Celles-ci peuvent être conçues pour traiter de données ponctuelles et/ou agrégées ainsi que pour détecter une tendance globale à l'agrégation ou pour localiser précisément les clusters. Les différentes variantes proposées dans la littérature afin de répondre à ces problématiques dans le cas prospectif sont présentées. Les méthodes permettant de répondre à l'objectif de mise en place d'un système de surveillance épidémiologique centré sur la détection de clusters spatiaux formulé en partie une, et compatibles avec ses contraintes sont isolées. Il en résulte que plusieurs peuvent être utilisées dans le cas de données agrégées pour une hypothèse de distribution de Poisson.

Dans la troisième partie, les capacités de détection des méthodes identifiées précédemment sont comparées à partir de données simulées proposées dans la littérature spécifiquement pour la réalisation de ce type d'études. Cinq méthodes pour données agrégées et compatibles avec une distribution de Poisson des cas observés sont évaluées sur la base de plusieurs critères. Les résultats obtenus permettent de choisir deux méthodes à privilégier pour la mise en oeuvre du système de surveillance et d'alerte défini à l'issue de la première partie.

La quatrième partie présente plusieurs applications épidémiologiques des méthodes de détection de clusters prospectives identifiées à la partie précédente. Trois contextes sont proposés : la détection d'épidémies, l'évaluation de leur impact sanitaire et le ciblage de territoires pour la mise en oeuvre d'actions de prévention et promotion de la santé. Dans le cas de la détection d'épidémies, deux situations ont été considérées : le suivi des épidémies de grippe saisonnière, et le suivi d'une pathologie émergente, la maladie à coronavirus de 2019. L'évaluation de l'impact des événements sanitaires exceptionnels a été réalisée à partir de la mortalité toutes causes, et la recherche de territoires cibles pour mener des actions de prévention portait sur les grossesses non désirées.

Enfin, une conclusion générale vient clore ce travail et dresser quelques perspectives.

2 La surveillance épidémiologique

2.1 Définition et concepts associés

Jusqu'à la deuxième guerre mondiale, le terme de surveillance n'était utilisé dans le cadre de la santé publique que pour désigner le suivi des sujets-contacts de patients atteints de pathologies contagieuses afin de déterminer s'ils avaient été contaminés ou non. C'est le développement des *Centers for Disease Control and Prevention* (CDC) aux Etats-Unis qui a entraîné l'élargissement du concept au domaine de l'épidémiologie. Le CDC a été fondé et installé à Atlanta (Géorgie) en 1946 pour prendre en charge les épidémies de paludisme qui sévissaient dans les états du Sud du pays depuis les années 30 et la Grande Dépression. Son action consistait principalement en un contrôle de la population de moustiques par la pulvérisation massive d'un insecticide, le DDT. En effet, la démarche épidémiologique de quantification et d'étude de l'épidémie était dans un premier temps jugé sans intérêt vu le très grand nombre de personnes atteintes dans l'ensemble du pays (6). C'est en 1950 qu'un programme national de surveillance du paludisme incluant une vérification systématique des cas est lancé et démontre son intérêt en aboutissant aux conclusions suivantes (6) :

- la plupart des cas déclarés étaient en réalité erronés ;
- la plupart des cas confirmés en laboratoire étaient importés ou anciens ;
- aucune épidémie localisée n'a pu être identifiée.

C'est suite à cet épisode que la surveillance épidémiologique s'est développée en étant étendue à d'autres pathologies et a été théorisée au CDC dès les années 50. Le terme de surveillance épidémiologique est apparu pour la première fois dans le cadre de l'activité de suivi de la poliomyélite en avril 1955 (6).

Langmuir, l'un des principaux artisans de son développement au CDC, en a donné la première définition (1) et la considérait comme :

> Le suivi de la distribution et des tendances de l'incidence via la collection systématique, la consolidation et l'évaluation de rapports de morbidité et de mortalité ou d'autres données pertinentes.

Il incluait aussi dans ce concept la diffusion des données et analyses réalisées dans ce cadre auprès des personnes y ayant contribué ou en ayant besoin.

Une définition plus récente mais proche est celle fournie par Thacker (1996) (7) :

> La surveillance en santé publique (aussi appelée surveillance épidémiologique) est le processus prospectif et systématique de collection, d'analyse et d'interprétation de données concernant des événements de santé spécifiques, et essentielles à la planification, l'implémentation et l'évaluation d'actions de santé publique, en relation étroite avec la diffusion auprès de ceux qui en ont besoin. Le dernier maillon de la chaîne de surveillance est l'application de ces données au contrôle et à la prévention des maladies humaines et des blessures.

A partir de ces définitions il est possible de faire ressortir quelques éléments essentiels d'un système de surveillance épidémiologique (2,8) :

- collecte de données de façon systématique et durable
- production d'indicateurs et d'analyses
- transmission de l'information aux décideurs pour servir la décision en santé publique

A côté de la surveillance épidémiologique on trouve d'autres concepts proches tels que la vigilance et la veille sanitaire.

La vigilance n'est pas distinguée de la surveillance chez les anglo-saxons, et en France elle a surtout une origine réglementaire et n'a pas bénéficié d'une définition épidémiologique claire. Elle correspond à un système de surveillance orienté vers les effets indésirables ou inattendus des produits ou des activités humaines et en particulier médicales (médicaments, produits dérivés du sang, cosmétiques, composés radioactifs, alimentation, etc.) (8). Elle repose sur la transmission d'événements perçus comme « anormaux » par les individus qui les signalent, et qui peuvent être des professionnels comme de simples particuliers. La difficulté réside alors dans l'hétérogénéité de la définition de l'anormalité parmi les producteurs de signaux, ainsi que dans l'exhaustivité des déclarations que l'on peut en attendre (8).

Le concept de veille sanitaire, plus récent, est à rapprocher du concept anglo-saxon d' *epidemic intelligence* (8). Il recouvre « l'ensemble des activités permettant d'identifier précocement des risques sanitaires potentiels, de vérifier leur véracité, de les évaluer et de réaliser les investigations nécessaires pour les documenter afin de permettre la prise de mesures de contrôle adéquates » (9). Les auteurs distinguent alors deux types de collection de données : l'une basée sur les indicateurs (de morbidité, de mortalité, etc.), en ce sens il s'agit d'une forme de surveillance non spécifique à rapprocher de la surveillance syndromique (voir section 2.4 Typologies), l'autre basée sur les événements, qu'il s'agisse de notification directe, de publication de cas atypiques, d'alertes médiatiques ou encore d'alertes survenant à l'étranger et susceptibles d'importation (9).

2.2 Objectifs

Les objectifs d'un système de surveillance épidémiologique sont de décrire l'état de santé de la population, d'alerter en cas d'anomalie ou de danger sanitaire, et d'évaluer les actions de santé publique (2). La description de l'état de santé d'une population et son évolution sont le préalable à toute démarche de surveillance. Elle suppose l'acquisition de données pertinentes concernant la morbidité, la mortalité, l'exposition à des risques sanitaires, ou tout événement pouvant impacter la santé ou être la manifestation d'une variation de l'état de santé des individus (ex : consommation de médicaments). Ces données sont ensuite utilisées pour la construction d'indicateurs spécifiques ou non, permettant la description des phénomènes de santé dans leur dimension temporelle et/ou spatiale.

Si les objectifs d'alerte et d'évaluation peuvent être assumés par un même système de surveillance, il est fréquent que chaque système soit spécialisé de façon à répondre à l'un de ces objectifs. Dès lors, il est possible de distinguer les systèmes orientés principalement

vers l'alerte, c'est-à-dire vers la détection d'épidémies ou de phénomènes émergents et potentiellement dangereux nécessitant une prise en charge rapide, de ceux orientés vers l'évaluation à moyen ou long terme de la tendance d'évolution de l'incidence d'une maladie ou de l'impact d'une action de santé publique (2).

Les systèmes d'alerte impliquent une rapidité de détection des anomalies, afin de pouvoir prendre une décision et mettre en oeuvre des actions correctives au plus vite. On va alors chercher à réduire le délai d'acquisition des données en privilégiant la simplicité en termes de critères de définition des cas, ou de volume d'informations recueillies pour chaque patient. Il est aussi possible de réutiliser les données qui sont produites de façon habituelle par le système de santé. Dans certains cas, il est également envisageable de se passer de l'exhaustivité en privilégiant une approche de type « sentinelles », où seul un échantillon d'établissements ou de professionnels de santé motivés contribue au recueil. A titre d'exemple, les réseaux de surveillance des pathologies infectieuses relèvent typiquement des systèmes d'alerte.

Les systèmes orientés sur l'évaluation des tendances temporelles ou géographiques nécessitent une définition plus précise et fiable des cas, ainsi qu'une certaine exhaustivité. La meilleure qualité et validité des indicateurs produits se fait alors au détriment de la rapidité d'acquisition, ce qui rend ces systèmes moins intéressants pour déclencher des alertes. Les registres de morbidité constituent un bon exemple de système d'évaluation.

2.3 Le processus de surveillance et d'alerte

Un système d'alerte cherche à identifier des événements inhabituels pouvant constituer une menace pour l'état de santé de la population. Il faut alors bien distinguer deux notions : le signal et l'alerte sanitaire. Le signal peut être défini comme « un événement de santé ou une situation d'exposition à un danger environnemental pouvant révéler une menace pour la santé publique » (4). L'alerte quant à elle correspond à « un signal validé pour lequel, après évaluation du risque, il a été considéré qu'il représente une menace pour la santé des populations et qui nécessite une réponse adaptée » (4). Afin de déterminer si un signal détecté par le système de surveillance doit donner lieu à une alerte, le processus de la surveillance sanitaire se décompose en plusieurs étapes (4) :

1. Collecte et analyse des données, construction d'indicateurs
2. Détection d'un signal
    1. Vérification du signal
    2. Validation du signal
    3. Évaluation de la menace
    4. Déclenchement d'une alerte
3. Gestion de l'alerte et prise de décision

Les sources de données sont variées et peuvent inclure des données médico-administratives (codage de l'activité hospitalière, certificats de décès, déclaration obligatoire, consommation de médicaments…), médicales (diagnostic clinique et/ou radiologique), biologiques (résultats de recherche d'agents pathogènes), ou encore

environnementales (allergènes, polluants atmosphériques, surveillance animale). Ces sources peuvent être combinées lorsque la définition des cas est composite, pour construire des indicateurs plus complexes ou pour combiner des indicateurs issus de plusieurs sources dans le processus de détection (ex : surveillance de la grippe en France qui combine fréquentation des urgences, recours à SOS médecin et données des centres de référence nationaux). Les indicateurs calculés sont généralement des indicateurs de morbidité (taux d'incidence, taux d'attaque) ou de mortalité.

Lorsque l'analyse des données conduit à la détection d'un signal (ex : détection d'une épidémie, d'un cluster spatial de cas), avant de déclencher une alerte sanitaire et donc des actions de santé publique, il faut le valider. La première étape consiste alors à vérifier le signal, c'est-à-dire à confirmer que le changement observé dans les données est bien dû à un changement réel dans le mode de survenue du processus surveillé. Il s'agit notamment de vérifier qu'il n'est pas lié à une erreur dans l'acquisition des données ou leur traitement, en vérifiant les critères de définition des cas ou en croisant les informations avec d'autres sources ou méthodes (4,9). La seconde étape consiste ensuite à valider le signal, c'est-à-dire à vérifier sa pertinence (le signal correspond-il bien à un problème de santé publique ?) (4,9).

Une fois qu'un signal a été validé, il convient d'évaluer la menace qu'il représente, ainsi que sa portée (locale, nationale ou internationale). Les critères communs aux signaux sanitaires et environnementaux qui peuvent signifier une menace de santé publique sont : le caractère inattendu de l'événement (épidémie d'une maladie connue, épidémie de cause inconnue, gravité supérieure à la normale, dissémination de produit toxique dans l'environnement, etc.), l'importance de l'impact potentiel pour la santé de la population (nombre de cas, mortalité importante, densité de population, contexte défavorable à la prise en charge) et le potentiel évolutif (risque d'extension à d'autres territoires, survenue dans une zone fortement exposée aux déplacements de population) (4). Concernant les signaux environnementaux, on peut y ajouter la toxicité potentielle des agents contaminants ainsi que les incertitudes relatives au contenu et à la nature exacte de la contamination.

Lorsque le signal est validé et la menace confirmée, l'alerte de santé publique doit être déclenchée afin d'informer les autorités compétentes. L'événement sort alors du champ de la surveillance épidémiologique pour entrer dans celui de la gestion du risque et de la décision publique.

2.4 Typologies

Les différents types de surveillance épidémiologique se distinguent principalement par le mode d'acquisition des données. La première distinction que l'on peut faire est entre les systèmes actifs ou passifs.

La surveillance passive se base sur la réutilisation de données, l'autorité en charge du programme n'intervenant pas directement dans leur collection. Il peut s'agir de systèmes utilisant les bases de données médico-administratives, les certificats de décès, le dossier médical électronique d'un établissement de soin ou encore les déclarations obligatoires

(même s'il y a contrainte légale, la charge de la déclaration des cas relève des individus et personne ne viendra rechercher les cas qui pourraient avoir été oubliés)(2).

Un système actif, au contraire, va impliquer une action directe de l'organisation qui en a la charge pour le recueil des informations. Par exemple, dans un registre de morbidité, des enquêteurs vont croiser plusieurs sources d'information (comme dans la surveillance passive), et solliciter directement les personnes ou organisations (hôpitaux, laboratoires d'analyse) susceptibles de disposer d'informations pertinentes. Ils vont aussi procéder à des vérifications systématiques pour confirmer ou infirmer les événements, les diagnostics posés, etc. Ces systèmes permettent donc potentiellement d'avoir un recueil de données plus exhaustif, plus précis et de meilleure qualité, mais il implique un coût humain et financier important, et l'acquisition des données est plus longue (2).

La surveillance peut aussi être conduite en réseau afin de favoriser la rapidité du recueil sur un vaste territoire (ex : RAISIN, le Réseau d'Alerte, d'Investigation et de Surveillance des Infections Nosocomiales en France). Ces systèmes sont alors généralement organisés autour d'un centre coordonnateur, chargé d'établir et valider les méthodes et les protocoles, de collecter les données des autres centres et d'en assurer le contrôle qualité, puis de les analyser (tâche qui peut être partagée entre plusieurs centres participants). Enfin, un retour d'information est réalisé auprès de tous les membres du réseau, ce qui peut contribuer, lorsque la surveillance porte sur des critères de qualité des pratiques, à produire un effet d'entraînement voire de compétition entre établissements pour l'amélioration des résultats. C'est par exemple ce qui s'est produit avec la plupart des réseaux de surveillance des infections nosocomiales en Europe et en Amérique du Nord (2).

On peut également distinguer les systèmes de surveillance selon le type d'indicateurs ou d'événements suivis pour opposer surveillance spécifique et non spécifique ou syndromique. La surveillance spécifique consiste à suivre à partir de critères précis des pathologies bien identifiées et connues pour présenter un risque pour la santé publique (ex : salmonellose) (4). La surveillance syndromique, ou non spécifique, a été définie par le CDC comme « une approche dans laquelle les équipes de santé publique, assistées par l'acquisition automatique de données et la génération de signaux statistiques, suit des indicateurs de santé en temps réel pour détecter des épidémies plus rapidement qu'avec les méthodes classiques » (10). Plus généralement, on peut dire que ce type de surveillance cherche à détecter des épidémies de pathologies non identifiées a priori à partir de données syndromiques aspécifiques (syndrome grippal, gastro-intestinal, rash cutané, etc.) ou de données non médicales pouvant manifester des problèmes de santé (ex : absentéisme à l'école) (4,11). Ils peuvent aussi se tourner vers les complications graves ou les décès inexpliqués dans le cas de la détection de pathologies émergentes dont les symptômes immédiats sont aspécifiques mais associées à un plus grand risque d'évolution défavorable (11). Contrairement à un système spécifique qui produit des données calibrées en fonction d'un objectif précis, la surveillance syndromique est construite à partir des données disponibles, sans cible précise (12).

Au-delà de ces différences, les systèmes de surveillance peuvent aussi être distingués en fonction de la temporalité du recueil (2). Dans certaines situations celui-ci est réalisé en continu, chaque cas identifié ou test positif étant transmis immédiatement ou très rapi-

dement au système. C'est le cas par exemple des maladies à déclaration obligatoire, ou encore des centres nationaux de référence. Une autre possibilité est la transmission groupée d'informations de façon périodique, comme c'est le cas par exemple des données hospitalières du Programme de Médicalisation des Systèmes d'Information (PMSI), avec une fréquence mensuelle.

Enfin, une dernière distinction peut se faire en fonction de la population cible (2). En effet, certains systèmes vont rechercher l'exhaustivité du recueil sur un territoire ou une population donnés. Il s'agit généralement de systèmes orientés vers l'évaluation et cherchant à produire les indicateurs les plus exacts possibles, comme les registres de morbidité par exemple. D'autres systèmes viseront plutôt un échantillon de population représentative pour des questions de faisabilité et de coût (ex : réseau Sentinelles de surveillance des pathologies infectieuses courantes en médecine générale).

2.5 Dispositifs de surveillance existants

De nombreux systèmes de surveillance épidémiologique ont été mis en place dans divers pays au fil du temps. Sans volonté d'exhaustivité, seront présentés ici succinctement quelques exemples représentatifs à l'international, puis en France.

2.5.1 A l'étranger

Aux Etats-Unis, les systèmes de surveillance épidémiologique sont gérés par des services de santé publique locaux sous la supervision du CDC d'Atlanta, qui assure la formation continue, l'émission de recommandations et la centralisation des données nationales. Ces systèmes peuvent être exclusivement locaux, tel celui de la ville de New-York, ou coordonnés à l'échelle fédérale comme les *Vital statistics* (certificats de décès) ou les *Surveys* (études spécifiques ciblées sur les comportements de santé ou certaines pathologies chroniques).

Le système RODS (pour *Real-time Outbreak and Disease Surveillance*), a démarré à l'Université de Pittsburgh en 1999 pour la surveillance des motifs de consultation aux urgences de son hôpital (12,13). Il suit 8 types de syndromes et est orienté vers la détection d'actes bioterroristes. Chaque séjour est classé dans l'une des catégories de syndrome à partir d'une analyse du texte libre fourni dans les notes médicales, puis les données peuvent être analysées par des méthodes de traitement du signal comme les *recursive least squares* (RLS) ou l'analyse par ondelettes. Il est aussi possible de détecter des événements anormaux par des méthodes de type *cumulative sum charts* (CUSUM). Le logiciel permet l'export de données ou de s'interfacer avec des modules statistiques externes tels que R ou SAS (*Statistical Analysis Software*). Le système a progressivement évolué pour devenir un logiciel libre distribué par l'université. Il a ensuite été adopté dans plusieurs états américains, comme l'Utah, ou grandes villes, comme Los Angeles, Houston ou Austin.

Le système de surveillance de la ville de New-York a été mis en place en deux temps à la suite des attentats du 11 septembre 2001 (12). Une première étape a été organisée en

urgence après les attentats afin de répondre à deux objectifs : la possibilité de détecter une attaque bioterroriste et l'évaluation de l'impact sanitaire des attentats. Cette première phase était réalisée avec un simple questionnaire papier comportant des données démographiques et de santé, à remplir pour chaque entrée aux urgences. Cette approche a vite montré ses limites face à la charge de travail qu'elle nécessitait pour les personnels hospitaliers (entraînant une baisse de la participation), comme pour le système de surveillance lui-même. Toutefois cette expérience a permis de faire évoluer son fonctionnement afin de passer à une nouvelle organisation dès novembre 2001. Cette fois-ci l'ensemble du processus était informatisé et automatisé de façon à gagner du temps et réduire les coûts, et le motif d'entrée était écrit en texte libre par les soignants, avant d'être classé dans les différentes catégories de syndromes étudiées par un algorithme informatique. Le système a également intégré à partir d'août 2002 le suivi des ventes d'un panel de médicaments dans la ville de New-York.

2.5.2 En France

En France, les systèmes de surveillance épidémiologique sont coordonnés par l'agence Santé Publique France, sous l'autorité du Ministère de la Santé et de sa Direction Générale de la Santé. Ces systèmes sont nombreux et regroupent aussi bien des registres de morbidité, que le système SurSaUD (Surveillance Sanitaire des Urgences et des Décès) de surveillance syndromique, le système RAISIN pour les infections associées aux soins ou le système de maladies à déclaration obligatoire.

Le système SurSaUD a été initié après la catastrophe sanitaire provoquée par la canicule de 2003, qui avait démontré l'incapacité du système national de surveillance sanitaire à détecter en temps voulu un événement sanitaire exceptionnel (12). Son objectif est d'organiser un suivi en routine assurant de pouvoir détecter rapidement des événements sanitaires importants, mais non définis précisément a priori. Il s'organise autour de la récupération des données de consultation de SOS médecins en ville (soit 55 associations locales couvrant principalement des territoires urbains), ainsi que du réseau OSCOUR (Organisation de la Surveillance Coordonnée des Urgences) de surveillance des consultations aux urgences. Ce dernier récolte des données démographiques et le motif de consultation codé via la CIM-10 (Classification Internationale des Maladies) pour chaque entrée aux urgences. Ces données sont centralisées quotidiennement par Santé Publique France, qui les analyse à différentes échelles géographiques, de façon globale et par regroupements de syndromes jugés pertinents. Un bulletin de synthèse hebdomadaire est produit à destination des professionnels de santé. Au 1er février 2015, environ 86 % des passages aux urgences étaient couverts par le système.

La surveillance des décès est de deux types. D'un côté l'état civil informatisé est centralisé quotidiennement par l'Insee (Institut national de la statistique et des études économiques), et regroupe des informations exclusivement d'ordre démographique : nom, prénom, date et lieu de naissance, date et lieu de décès (12). Si l'état civil est informatisé, la déclaration du décès elle reste majoritairement réalisée par formulaire papier, bien que les certificats électroniques progressent régulièrement. Ces données d'état civil sont publiées en libre accès sur la plate-forme de données ouvertes du gouvernement depuis 2019. D'un autre

côté, les causes médicales de décès anonymisées sont transmises par les médecins à l'Inserm (Institut national de la santé et de la recherche médicale), qui centralise ces données dans l'unité CépiDC (Centre d'épidémiologie sur les causes médicales de décès) pour les analyser. Ces données font partie intégrante du Système National des Données de Santé (SNDS) et ne sont accessibles que pour des projets de recherches.

2.6 Problématiques actuelles

Ces dernières années, en France, plusieurs polémiques sanitaires se sont développées autour de cas d'agrégats localisés d'événements de santé, comme des malformations ou des cancers, évoqués par des lanceurs d'alerte, qu'ils soient professionnels de santé ou particuliers. L'affaire des agénésies transverses du membre supérieur dans l'Ain constitue l'un des exemples les plus médiatisés (3).

Dès lors se pose la question de savoir si ces agrégats perçus sont simplement dus au hasard, ou s'ils témoignent d'un phénomène réel nécessitant d'avantage d'investigations. En effet, un biais bien connu est l'attente excessive d'étalement (ou « effet rateau »), qui est dû au fait que notre esprit n'a pas une représentation exacte de ce qu'est une répartition aléatoire de points dans l'espace, et s'attend donc à ce qu'ils soient répartis plus régulièrement que ne le ferait le hasard (14). Des méthodes statistiques de détection d'agrégats pourraient permettre de répondre objectivement à cette interrogation.

Cette actualité témoigne d'un besoin, en France, pour l'étude systématique et la détection de territoires nécessitant une intervention de santé publique. Ces besoins d'intervention peuvent porter sur différentes thématiques :

- une épidémie de maladie transmissible connue (comme la grippe) ou émergente (comme la maladie à coronavirus de 2019) : l'objectif est alors la lutte contre la diffusion de l'épidémie
- l'apparition d'un facteur de risque environnemental au sens large (que l'on parle d'environnement naturel ou humain) : il s'agit dans ce cas d'identifier ce(s) facteur(s), puis de protéger la population exposée
- le ciblage d'actions de prévention vers les territoires en ayant le plus besoin, en situation de ressources rares, ainsi que le suivi et l'évaluation de ces interventions

Dans le cadre de ce travail de thèse, on se propose de faire une étude de faisabilité de la mise en place d'un système de surveillance épidémiologique permettant la détection de territoires nécessitant une intervention de santé publique. Ce système serait passif et orienté vers l'alerte, et se placerait dans le cadre de la surveillance syndromique, au sens de la réutilisation de données de santé déjà acquises en routine.

3 Les méthodes de surveillance

Les méthodes de surveillance épidémiologique se sont concentrées historiquement sur la surveillance de l'évolution dans le temps des problèmes de santé à l'échelle de territoires entiers, où séparément à celle de leurs sous-unités administratives, comme les régions. Toutefois cette approche ne prend pas en compte la dimension spatiale des données acquises et la proximité géographique des cas observés, ce qui peut nuire aux capacités de détection d'événements sanitaires lorsqu'ils touchent plusieurs régions contiguës (15,16).

Dans le cadre de la recherche de territoires d'intérêt, cette section se concentrera principalement sur les méthodes de surveillance prenant en compte la dimension spatiale des séries temporelles observées.

3.1 Suivi temporel

Ne seront abordées ici que les éléments utiles pour l'introduction des méthodes de surveillance spatio-temporelle qui suivront. Pour une vue plus détaillée des méthodes de surveillance temporelle il est possible de se rapporter à la revue d'Unkel et al. (15).

3.1.1 Méthode de Shewhart

La méthode de Shewhart est issue du contrôle statistique de processus industriel. Il s'agissait alors de vérifier qu'un paramètre quantitatif mesuré sur la production industrielle restait conforme à une norme pré-établie au cours du temps. Pour cela, Shewhart a proposé de répéter régulièrement dans le temps un test statistique afin de comparer la valeur obtenue sur un échantillon de production à sa norme (17).

Sous l'hypothèse nulle la variable aléatoire étudiée X (par exemple le diamètre d'une pièce mécanique) est supposée suivre une loi normale de moyenne μ_0 et de variance σ_0^2 connues, il s'agit des normes attendues. Sous l'hypothèse alternative la production ne correspond plus à ces normes, et la variable étudiée présente une moyenne μ_1 différente de μ_0. Lorsque l'on s'intéresse à une situation de santé publique, la variable aléatoire devient alors généralement un nombre d'événements de santé (par exemple le diagnostic d'une pathologie), et les valeurs attendues pour sa moyenne et sa variance sous l'hypothèse nulle peuvent être estimées à partir de données historiques.

Pour tester ces hypothèses, une statistique de test z_t est calculée à chaque temps de l'analyse à partir de la valeur x_t de X au temps t de telle sorte que :

$$z_t = \frac{x_t - \mu_0}{\sigma_0}$$

Un seuil de significativité h est déterminé à partir de la distribution de la loi normale centrée réduite et du risque de première espèce α fixé à l'avance, soit par exemple $h = 1.96$ si l'on souhaite garantir $\alpha = 0.05$ dans le cadre d'un test bilatéral. L'hypothèse nulle est rejetée au temps t si $|z_t| > h$.

Cette méthode est généralement présentée sous la forme d'un graphique représentant le temps en abscisse et la statistique de test en ordonnée. Les seuils de significativité sont représentés par des lignes horizontales d'ordonnées h et $-h$ formant des limites supérieure et inférieure, lorsque la statistique de test dépasse l'une de ces limites, l'hypothèse nulle est rejetée.

3.1.2 Cumulative sum charts (CUSUM)

3.1.2.1 Notation classique La méthode de Shewhart présentée précédemment reposait sur un principe simple : mesurer régulièrement le paramètre étudié, et vérifier qu'il ne s'éloignait pas de la moyenne attendue de plus de h écarts-types. Mais dans ce cas, seule la dernière valeur mesurée était utilisée, il y avait donc une perte de puissance pour identifier un écart à la norme attendue. C'est pourquoi Page (1954) (18) a proposé une méthode cumulative permettant d'exploiter l'information contenue dans plusieurs mesures successives au cours du temps lors du test statistique : la *cumulative sum charts* (CUSUM). Le calcul de la statistique de test repose sur la sommation dans le temps d'un score qui, en moyenne, est négatif lorsque la valeur mesurée respecte la norme attendue, et positif sinon. Pour cela le score adopté correspond généralement aux écarts constatés entre la valeur observée et une valeur de référence, qui dépend de la valeur attendue et de l'amplitude de variation que l'on souhaite pouvoir détecter.

Sous l'hypothèse nulle (H0), on suppose que la variable aléatoire étudiée x_t au temps t suit une loi normale de moyenne μ_t, le nombre moyen d'événements constatés sur une période historique dénuée d'épidémie, et de variance σ_t^2, estimée sur la même période. Les valeurs successives de x_t sont supposées indépendantes. Sous l'hypothèse alternative (H1), il existe un moment noté τ où le risque de survenue de l'événement est augmenté de $r > 0$, et x_t ($t \geq \tau$) suit alors une loi normale de moyenne $\mu_t + r$ et de variance σ_t^2. La statistique de somme cumulée gaussienne unilatérale (hypothèse d'augmentation du risque) au temps t est alors définie par :

$$S_t = \max(0, S_{t-1} + \frac{x_t - \mu_t}{\sigma_t} - k) = \max(0, S_{t-1} + z_t - k)$$

avec $S_0 = 0$, $z_t = \frac{x_t - \mu_t}{\sigma_t}$ un z-score calculé à chaque temps t, et k une valeur de référence, donnée en nombre d'écarts-types, et choisie entre 0 et l'amplitude de variation minimale que l'on souhaite pouvoir détecter (18,19). Cette amplitude est généralement fixée à un écart-type, et la valeur de k à 0.5, étant entendu que la valeur de 1/2 de l'amplitude choisie est censée minimiser le délai de détection pour un risque de faux positif fixé.

L'hypothèse nulle est rejetée si la statistique S_t dépasse le seuil critique h, fixé à l'avance.

Dans le cadre des CUSUM, le taux de faux positifs est contrôlé par le biais de l' *average run length* sous l'hypothèse nulle (ARL_0), qui correspond dans ce cas au nombre moyen d'observations nécessaire pour obtenir une alarme sous H0, sachant qu'un test est réalisé

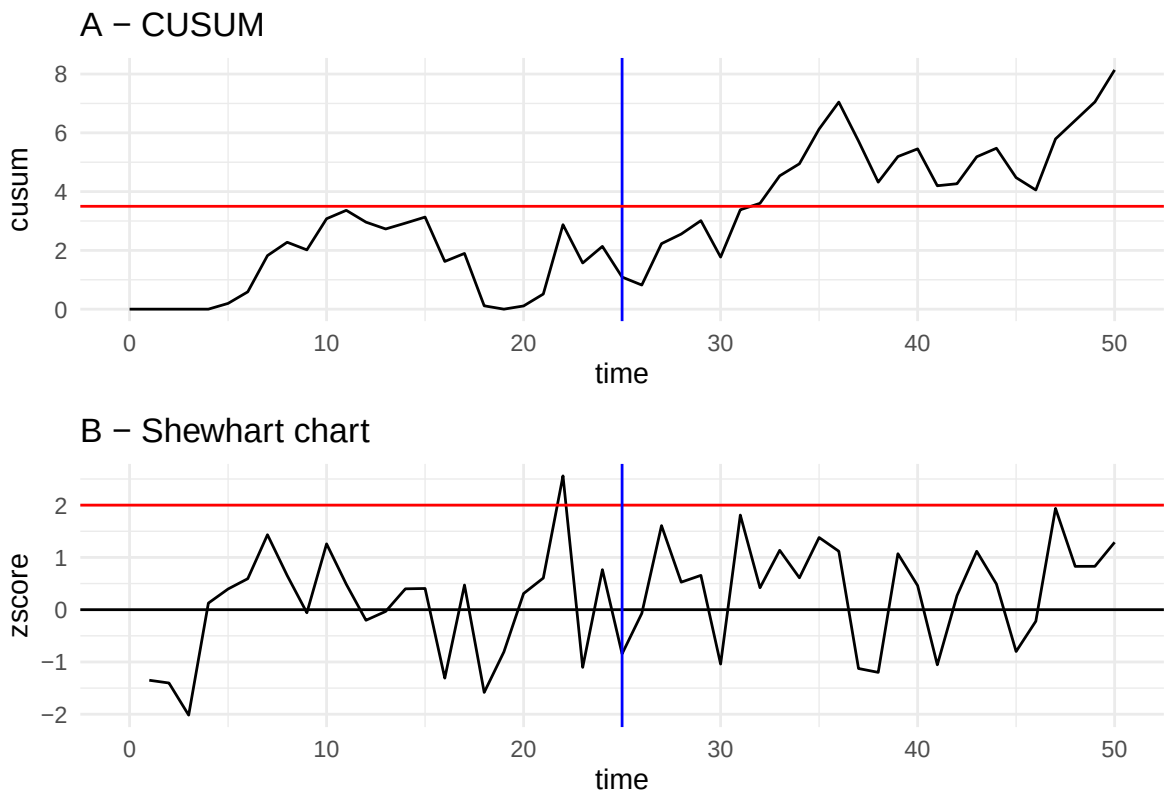

FIG. 1 : Illustration du diagramme de Shewhart (B) et de la CUSUM correspondante (A). La ligne bleue indique le moment du passage de la moyenne du Z-score de 0 à 0,5. La ligne rouge indique le seuil d'alerte pour un risque alpha unilatéral de 2,5 % (B) ou un ARL de 40 (A).

à chaque itération. Le seuil h peut être fixé à partir de la formule d'approximation donnée par Rogerson (2006) (20) à partir de Siegmund (1985) (21), valable pour $k = \frac{1}{2}$ et une hypothèse de distribution selon la loi normale :

$$ARL_0 \approx 2(\exp(h + 1.166) - h - 2.166)$$

et qui peut être inversée (20) pour donner :

$$h \approx \left(\frac{ARL_0 + 4}{ARL_0 + 2}\right) \log\left(\frac{ARL_0}{2} + 1\right) - 1.166$$

Cette méthode a été largement utilisée dans le domaine de la surveillance épidémiologique temporelle (15) sous sa forme gaussienne ou dans sa version adaptée à la distribution de Poisson (19) lorsque le nombre de cas est trop faible pour utiliser une approximation normale. La moyenne μ_t sous l'hypothèse nulle est alors égale à la variance et peut être estimée à partir de données historiques antérieures et dénuées d'épidémie.

3.1.2.2 Notation alternative Il a été dit précédemment que la CUSUM repose sur la sommation dans le temps d'un score qui est en moyenne négatif lorsque la valeur mesurée respecte la norme attendue, et positif sinon. Plutôt que de calculer ce score par une différence entre la valeur observée et une valeur de référence, il a été proposé d'utiliser un rapport de vraisemblance.

Pour rappel, sous l'hypothèse nulle la variable étudiée x_t suivait une loi normale de moyenne μ_t et de variance σ_t^2. Sous l'hypothèse alternative, il existait un temps τ à partir duquel la moyenne de la loi normale était augmentée à $\mu_t + r$ avec $r > 0$ et $t \geq \tau$ tandis que la variance était inchangée. Si x_t mesure un nombre d'événements de santé (par exemple un diagnostic de cancer) survenus au temps t, cette augmentation traduit un accroissement du risque de survenue de l'événement.

On note $LR_\tau(t)$ le rapport de la vraisemblance des données observées sous l'hypothèse alternative sur la vraisemblance sous l'hypothèse nulle, calculée au temps t en supposant que l'augmentation du risque est survenue au temps $\tau \leq t$. La vraisemblance sous l'hypothèse alternative peut être calculée en fixant $r = r^*$ avec r^* l'ampleur minimale de l'augmentation que l'on souhaite pouvoir détecter.

Il est alors possible d'utiliser comme score pour le calcul de la CUSUM le logarithme de ce rapport de vraisemblance en supposant que l'augmentation du risque est survenue au temps t de l'analyse. En effet, dans ce cas si l'hypothèse nulle est vraie on s'attend à ce que la valeur de $LR_t(t)$ soit inférieure à 1 et donc que son logarithme soit négatif, tandis que si l'hypothèse alternative est vraie on s'attend à ce que $LR_t(t)$ soit supérieure à 1 et que son logarithme soit donc positif. La statistique de somme cumulée unilatérale (seule une augmentation du risque est recherchée) est alors :

$$S_t = \max(0, S_{t-1} + \ln(LR_t(t)))$$

Dans une perspective de santé publique, il est aussi intéressant de savoir quand l'augmentation du risque est survenue, donc d'estimer la valeur de τ. Cette estimation peut être réalisée par maximum de vraisemblance, en recherchant la valeur $\widehat{\tau}$ qui maximise le rapport de vraisemblance $LR_\tau(t)$. Il a été suggéré par Page (18), puis démontré par Sonesson (22) que cette formulation de la CUSUM donne précisément le maximum de vraisemblance de $LR_\tau(t)$ en fonction de τ (avec $1 \leq \tau \leq t$), ce qui donne :

$$S_t = \max(0, S_{t-1} + \ln(LR_t(t))) = \max_\tau LR_\tau(t)$$

La valeur de $\widehat{\tau}$ correspond alors au dernier temps t pour lequel S_t avait une valeur nulle (18).

3.1.3 Statistique de Shiryaev-Roberts

La méthode de Shiryaev-Roberts a été proposée également dans le cadre du contrôle statistique de processus (23–25). Elle repose sur un principe similaire à celui de la CUSUM.

Sous l'hypothèse nulle on suppose que la variable étudiée x_t à chaque temps t suit une loi de probabilité donnée, par exemple une loi de Poisson de paramètre λ_0 pour le nombre de cas incidents d'une pathologie. Les x_t successifs ne sont pas nécessairement indépendants. Sous l'hypothèse alternative, il existe un temps τ à partir duquel le risque de survenue de l'événement est augmenté et x_t suit alors une loi de Poisson de paramètre $\lambda_1 > \lambda_0$.

La fonction de vraisemblance sous l'hypothèse alternative au temps t est notée $f_\tau(x_1, \ldots, x_t)$. Sous l'hypothèse nulle, aucune augmentation du risque de l'événement n'est survenue, ce qui revient à considérer que l'augmentation survient à un temps infini, soit $\tau = \infty$. La fonction de vraisemblance sous l'hypothèse nulle au temps t est donc notée $f_\infty(x_1, \ldots, x_t)$, et le rapport de vraisemblance est :

$$LR_\tau(t) = \frac{f_\tau(x_1, \ldots, x_t)}{f_\infty(x_1, \ldots, x_t)}$$

Contrairement à la CUSUM, qui calcule le maximum de ce rapport de vraisemblance en fonction de τ, la statistique de Shiryaev-Roberts R_t est la somme des rapports de vraisemblance pour toutes les valeurs possibles de τ, soit :

$$R_t = \sum_{\tau=1}^{t} LR_\tau(t) = \sum_{\tau=1}^{t} \frac{f_\tau(x_1, \ldots, x_t)}{f_\infty(x_1, \ldots, x_t)}$$

L'hypothèse nulle est rejetée si la statistique R_t dépasse le seuil critique A, fixé à l'avance.

Dans le cadre de cette statistique, comme pour les CUSUM, le taux de faux positifs est contrôlé par le biais de l' *average run length* sous l'hypothèse nulle (ARL_0), défini précédemment. Il a été montré que A est approximativement égal à l' ARL_0, ce qui permet de prendre comme seuil d'alerte l'ARL_0 que l'on souhaite garantir (23).

3.2 Détection de clusters spatiaux

3.2.1 Les types de données spatiales

Lorsque l'on s'intéresse à des événements de santé localisés dans l'espace et le temps, on peut être confronté à différents types de données spatiales (26). De façon générale, les données spatio-temporelles sont représentées par $X = \{X_{d,t}, d \in D, t \in T\}$, un processus aléatoire indexé par un ensemble spatial D et un ensemble temporel $T \subseteq \mathbb{R}^+$ et à valeurs dans un espace d'états E. L'ensemble spatial D peut être constitué de 2 dimensions ($D \subseteq \mathbb{R}^2$) ou 3 dimensions ($D \subseteq \mathbb{R}^3$), et chacun des sites d de l'ensemble peut avoir des coordonnées fixées ou aléatoires en fonction du modèle considéré. L'espace d'états E peut être de différents types en fonction de la nature de la variable étudiée :

- binaire : $E = \{0, 1\}$;
- catégorielle : $E = \{a_0, a_1, \dots, a_k\}$;
- quantitative discrète : $E \subseteq \mathbb{N}$ (champ poissonien) ;
- quantitative continue :
  - $E \subseteq \mathbb{R}^p$ (champ gaussien) ;
  - $E \subseteq \mathbb{R}^+$ (champ exponentiel).

Des données spatio-temporelles de santé correspondent donc à n réalisations $\{x_{d_1,t_1}, \dots, x_{d_n,t_n}\}$ d'une variable aléatoire X ayant pour coordonnées spatiales $\{d_1, \dots, d_n\}$, et pour coordonnées temporelles $\{t_1, \dots, t_n\}$. Les différents types de données spatiales se distinguent par la nature des ensembles D et E. Les trois principaux types de données spatiales sont les données géostatistiques, les données latticielles et les données ponctuelles.

Les données géostatistiques correspondent au cas où l'on mesure une variable quantitative continue ($E \subseteq \mathbb{R}^p$ avec $p \geq 1$) à partir de n sites fixés $\{d_1, \dots, d_n\} \in D$, dans un sous-espace D continu fixé de $\mathbb{R}^q$ ($q \geq 2$). Autrement dit, la variable quantitative étudiée pourrait théoriquement être mesurée à n'importe quelle localisation du sous-espace D (continu), mais en pratique elle ne l'est qu'en certaines localisations d_i fixées. Cette situation correspond par exemple à celle de données météorologiques, comme des températures, qui sont mesurées à partir de stations de mesures pré-déterminées, mais qui en réalité prennent une valeur à n'importe quel point de l'espace continu.

Les données latticielles correspondent à la situation où une variable (par exemple un nombre d'événements) est mesurée ou agrégée à l'échelle d'une unité spatiale, comme une commune ou un département par exemple. Dans ce cas, D est un ensemble discret fixé structuré par un graphe de voisinage $G = \{D, A\}$, où D est l'ensemble des

sommets du graphe et A l'ensemble de ses arêtes. D représente l'ensemble des unités spatiales choisies pour l'agrégation des données (ex : commune), et contient M éléments d_i. Le graphe G est muni d'une matrice d'adjacence U de taille $M \times M$, et dont chaque élément $u_{i,j}$ ($i, j \in \{1, ..., M\}$) vaut 1 si les sommets d_i et d_j sont reliés par une arête (soit $(i, j) \in A$), et 0 sinon. Les arêtes du graphe de voisinage indiquent le fait que deux unités spatiales soient contiguës, c'est-à-dire que seules les unités voisines sont reliées entre elles.

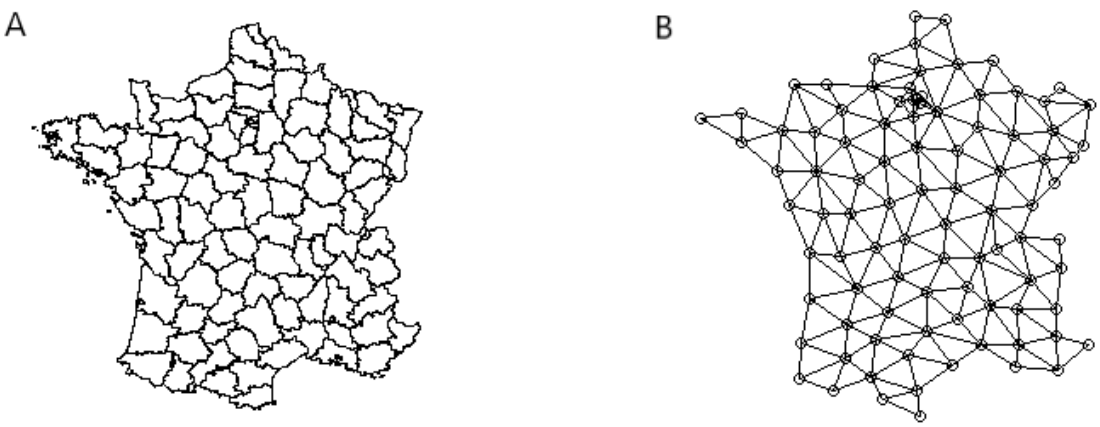

FIG. 2 : Carte des départements français en A, et leur graphe de voisinage en B

Les données ponctuelles correspondent au cas où les événements sont localisés individuellement, c'est-à-dire que pour chaque événement particulier on connait le lieu et la date précise où il a eu lieu. Cette situation est modélisée par la notion de processus spatial ponctuel. La variable d'intérêt est ici la localisation des événements et, contrairement aux deux situations précédentes, les coordonnées des sites d_i sont aléatoires et déterminées par $x = \{x_1, ..., x_n\}$. Dans cette situation, le nombre total de sites $n(x)$ est lui-même aléatoire. La réalisation du processus spatial ponctuel X est $x = \{x_1, ..., x_n\}$, et est appelée semis de points. Le processus peut être marqué par association à chaque point d_i d'une marque m quantitative (ex : âge) ou qualitative (ex : catégorie socio-professionnelle).

Si certaines méthodes de détection d'agrégats d'événements (clusters) peuvent être adaptées à ces deux derniers types de données, la plupart ne sont présentées que pour l'étude de l'un des deux.

Deux approches principales ont été proposées pour le suivi spatio-temporel de données ponctuelles. L'une se base sur le suivi dans le temps d'un indicateur de tendance à l'agrégation spatio-temporelle par l'intermédiaire de la méthode des *cumulative sum charts* (CUSUM). L'autre, d'une façon similaire, étend au domaine spatial une méthode apparentée à la CUSUM : la méthode de Shiryaev-Roberts.

Dans le cadre de l'analyse de données agrégées, on distinguera une méthode dédiée au suivi de la tendance globale à l'agrégation, mais ne permettant pas de localiser l'éventuel cluster qui en serait à l'origine, des méthodes permettant de localiser le cluster détecté. Ces dernières partagent toutes un même mécanisme de fonctionnement, impliquant un processus de balayage de l'espace et une phase d'inférence permettant de comparer les effectifs d'événements observés aux effectifs attendus. Pour plus de clarté, elles ont donc été regroupées en présentant les différentes modalités possible à chaque étape du processus, suivies d'une présentation des méthodes de calcul des cas attendus.

3.2.2 La cumulative sum charts (CUSUM) de Knox

La première méthode proposée dans le cas de données ponctuelles suit l'évolution dans le temps d'un indicateur de tendance globale à l'agrégation spatio-temporelle : la statistique de Knox, et a été proposée par Rogerson (2001) (27). La statistique de Knox permet de tester de façon rétrospective l'interaction spatio-temporelle dans un processus ponctuel, c'est-à-dire la propension de deux points proches dans l'espace à être aussi proches dans le temps.

Sous l'hypothèse nulle, les cas sont répartis de façon aléatoire dans l'espace et dans le temps selon un processus ponctuel. Sous l'hypothèse alternative, il existe un cluster dans lequel les points sont plus proches dans l'espace et dans le temps qu'ils ne le seraient s'ils étaient répartis de façon aléatoire.

Soient n_s le nombre de paires de points qui sont proches dans l'espace (c'est-à-dire séparés d'une distance inférieure à un seuil s préfixé), et n_t le nombre de paires de points qui sont proches dans le temps (c'est-à-dire séparés d'une distance inférieure à un seuil t préfixé). Les valeurs seuils sont définies par l'utilisateur en fonction de la taille prévisible des clusters à détecter (compte tenu des caractéristiques de la pathologie suivie) (27,28). Plusieurs valeurs peuvent être testées, comme le fait Rogerson dans son article (27). La statistique de Knox est le nombre de paires de points $n_{s,t}$ qui sont à la fois proches dans l'espace et dans le temps. La variable aléatoire associée à cette quantité est $N_{s,t}$. La distribution sous l'hypothèse nulle de cette variable est, selon les études, approchée par une loi de Poisson, une loi normale ou estimée par méthode de Monte Carlo avec permutations des temps de survenue.

La statistique de Knox permet donc de réaliser un test évaluant une tendance globale au clustering. La méthode proposée par Rogerson consiste à créer une version locale de cette statistique afin de s'en servir pour étudier l'impact de chaque nouvel événement sur la tendance à l'agrégation. Cette statistique locale peut être calculée pour chaque point i ($1 \leq i \leq n$), représentant une observation localisée dans le temps et dans l'espace. Elle évalue alors si les points du processus étudié ont tendance à être à la fois plus proches dans l'espace et plus proches dans le temps du point i considéré, qu'on ne l'attendrait sous l'hypothèse nulle, les hypothèses restant les mêmes que précédemment.

Soient $n_s(i)$ le nombre points qui sont proches de i dans l'espace, $n_t(i)$ le nombre de points qui sont proches de i dans le temps et $n_{s,t}(i)$ le nombre points qui en sont proches

à la fois dans le temps et dans l'espace. La statistique de Knox locale observée est $n_{s,t}(i)$ et la variable aléatoire associée est $N_{s,t}(i)$.

La distribution de cette statistique peut être obtenue par des permutations des dates de survenue des événements étudiés. En supposant que ces permutations sont équiprobables, Rogerson propose une approximation normale de cette distribution et donne l'espérance $E(N_{s,t}(i))$ et la variance $V(N_{s,t}(i))$ associées. La statistique de test est alors un z-score qui peut être comparé à la distribution de la loi normale centrée réduite. Le z-score est donné par :

$$z_i = \frac{n_{s,t}(i) - E(N_{s,t}(i)) - 0.5}{\sqrt{V(N_{s,t}(i))}}$$

La statistique de Knox locale permet, lorsque le test global est significatif, de localiser les observations qui y contribuent.

Pour réaliser un suivi dans le temps de l'évolution de cette statistique, il utilise la méthode des *cumulative sum charts* (CUSUM).

Sous l'hypothèse nulle, les points sont répartis de façon aléatoire dans l'espace et dans le temps. Sous l'hypothèse alternative, il existe un temps τ à partir duquel apparait un cluster dans lequel les points sont plus proches dans l'espace et dans le temps qu'ils ne le seraient s'ils étaient répartis au hasard.

Dans le cas prospectif on veut savoir si la valeur de la statistique de Knox observée après chaque nouveau cas i (notée K_i) est différente de celle qui serait observée sous l'hypothèse nulle, conditionnellement à sa valeur après l'observation i-1 et aux valeurs observées $n_s(i)$ et $n_t^j(i)$ ($1 \leq j \leq i$). Ce suivi peut être réalisé à partir du z-score suivant :

$$z_i = \frac{K_i - E\{K_i | K_{i-1}, n_s(i), n_t(i)\} - 0.5}{\sqrt{V\{K_i | K_{i-1}, n_s(i), n_t(i)\}}}$$

L'information contenue dans ce z-score étant uniquement due à l'effet spécifique de la dernière observation i, Rogerson démontre son identité avec le z-score associé à la statistique de Knox locale, avec $n_{s,t}(i) = K_i - K_{i-1}$, et donne les espérance et variance correspondantes.

La statistique de somme cumulative (CUSUM) S_i est alors définie par :

$$S_i = \max(0, S_{i-1} + z_i - k)$$

avec $S_0 = 0$ et k une valeur de référence, donnée en nombre d'écarts-types, et choisie entre 0 et l'amplitude de variation minimale que l'on souhaite pouvoir détecter. Cette amplitude est généralement fixée à un écart-type, et la valeur de k à 0.5, étant entendu que

la valeur de 1/2 de l'amplitude choisie est censée minimiser le délai de détection pour un risque de faux positif fixé.

L'hypothèse nulle est rejetée si la statistique S_i dépasse le seuil critique h, fixé à l'avance. Le taux de faux positifs est contrôlé par le biais de l'average run length sous l'hypothèse nulle (ARL_0), que Rogerson propose de fixer le seuil h à partir de la formule d'approximation donnée par Siegmund (1985), valable pour $k = \frac{1}{2}$ (cf section 3.1.2 *Cumulative sum charts*).

La méthode de Rogerson a été critiquée par Marshall (2007) (29), qui a montré que, contrairement à l'ARL_0 approché à partir de la formule donnée par Rogerson, l'ARL_0 réel était fortement influencé par les valeurs des seuils de distance fixées par l'utilisateur, la densité de population et la forme du cluster. C'est pour répondre à ces critiques que Piroutek et al. (28) ont proposé une correction de cette méthode.

Ces derniers expliquent que bien que Rogerson présente sa méthode comme prospective, en réalité le mode de calcul de la statistique locale prend en compte l'ensemble des points proches du point étudié (en fonction des seuils choisis par l'utilisateur), qu'ils soient apparus avant ou après lui. Cette situation est donc par définition rétrospective. Piroutek et al. proposent alors des modifications à la méthode de Rogerson de façon à la rendre prospective.

Les hypothèses nulle et alternative sont identiques à celles de Rogerson. La modification principale proposée par les auteurs consiste à ne considérer pour le calcul des $n(i)$ que les observations antérieures à i, là où la méthode de Rogerson permet le compte d'événements postérieurs au point étudié. Ce changement entraine une mise à jour des formules de calcul de l'espérance et de la variance de la variable aléatoire :

$$E(N_{s,t}(i)) = \frac{n_s(i)n_t(i)}{i-1}$$

$$V(N_{s,t}(i)) = \frac{n_s(i)n_t(i)(i-1-n_s(i))(i-1-n_t(i))}{(i-2)(i-1)^2}$$

La statistique de somme cumulée étant ensuite calculée de la même façon que précédemment.

L'autre modification proposée par Piroutek et al. concerne le choix du seuil d'alerte h. En effet la formule de calcul proposée par Rogerson se base sur une hypothèse de normalité qui peut être fausse si le nombre de cas attendu est faible. Les auteurs proposent donc une méthode de détermination empirique du seuil se basant sur des simulations par permutations.

Cette méthode requiert un nombre $m+n$ d'observations antérieures. Les m premières observations servent à stabiliser le score local, car en l'absence d'observations antérieures il sera proche de zéro ce qui ne reflète pas sa distribution réelle. Les n observations suivantes sont utilisées pour le calcul des scores. Les expériences des auteurs les conduisent

à conseiller de garantir $m \geq 200$ pour un choix de seuils s et t tels que $E(N_{s,t}(i)) \geq 3$. On génère R jeux de données simulés par permutation des $m+n$ dates de survenue des événements. Pour chaque permutation j ($1 \leq j \leq R$) on calcule les statistiques de sommes cumulées S_i^j correspondantes.

On se fixe un risque α d'obtention d'un faux positif parmi un nombre d d'observations successives. Pour chaque observation i ($m \leq i \leq m+n$) de chaque permutation j, on calcule $MS_i^j = \max_t\{S_t^j, i+1 \leq t \leq i+d\}$, le maximum des valeurs de la CUSUM parmi les d observations suivant le cas i dans la permutation j. Pour chaque i ($m \leq i \leq m+n$), on recherche le percentile $1-\alpha$ des MS_i^j ($1 \leq j \leq R$), qui donne le h_i permettant de garantir α sur d observations en partant du temps i. A partir de ces valeurs on peut alors calculer le seuil d'alerte global du système :

$$h = \frac{\sum_{i=1}^{n} h_i}{n}$$

3.2.3 Méthode de Shiryaev-Roberts spatiale

La méthode de Shiryaev-Roberts est apparentée à la CUSUM, et suit dans le temps une somme de rapports de vraisemblance. Elle est à l'origine uniquement dédiée au suivi temporel, mais Assunção et Correa (2009) (30) ont proposé de l'étendre au cas spatio-temporel pour données ponctuelles. Pour cela, ils introduisent la notion de balayage de l'espace à 3 dimensions (2 dimensions d'espace pour une dimension de temps) avec une fenêtre de forme cylindrique. Cette fenêtre permet de balayer les clusters candidats potentiels pour calculer la statistique.

Soit N un processus ponctuel de Poisson défini sur un espace $D \subseteq \mathbb{R}^2$ et un intervalle temporel $T \subseteq \mathbb{R}^+$. Les événements $(d_i, t_i) = (x_i, y_i, t_i)$ sont indexés par $i \in \mathbb{N}$ avec $t_1 < t_2 < ... < t_n$ et n le nombre d'événements. La fonction d'intensité du processus est $\lambda(x, y, t)$. Si $N(z)$ est le nombre d'événements dans l'ensemble $z \subseteq D \times T$, $N(z)$ suit une loi de Poisson de moyenne $\mu(z)$, l'intégrale de la fonction d'intensité sur z :

$$\mu(z) = \int_z \lambda(x, y, t) \mathrm{d}x \mathrm{d}y \mathrm{d}t$$

Soit $\mu = \mu(D \times T)$ l'espérance du nombre d'événements sur l'ensemble du territoire étudié pendant la durée du suivi. Sous l'hypothèse nulle (H0) d'absence de cluster on a $\lambda(x, y, t) = \mu\lambda_D(x, y)\lambda_T(t)$, avec $\lambda_D(x, y)$ et $\lambda_T(x, y)$ les densités marginales spatiale et temporelle respectivement.

Sous l'hypothèse alternative (H1), on a :

$$\lambda(x, y, t) = \mu\lambda_D(x, y)\lambda_T(t)(1 + \epsilon I_z(x, y, t))$$

avec $\epsilon > 0$ le risque relatif anticipé dans le cluster z (fixé par l'utilisateur) et $I_z(x, y, t)$ une fonction indicatrice valant 1 quand $(x, y, t) \in z$ et 0 quand $(x, y, t) \notin z$.

Soit $z_{k,n}$ une fenêtre de balayage spatio-temporel cylindrique de rayon fixe ρ représentant un cluster candidat. Chaque cylindre débute à la date t_k de l'un des événements k observés, et est centré sur le site d_k correspondant. Chaque fenêtre se termine au temps de la dernière observation disponible t_n.

On note L_∞ la vraisemblance associée au processus sous l'hypothèse nulle, et L_k la vraisemblance sous l'hypothèse alternative. Le rapport de vraisemblance est alors :

$$\Lambda_{k,n} = \frac{L_k}{L_\infty} = (1+\epsilon)^{N(z_{k,n})} \exp(-\epsilon\mu(z_{k,n}))$$

L'espérance $\mu(z_{k,n})$ du nombre d'événements dans la fenêtre z peut-être estimée par le produit du nombre total d'événements survenus à une distance inférieure à ρ du point (d_k, t_k) par le nombre total d'événements survenus dans l'intervalle de temps $[t_k, t_n]$, divisé par le nombre total d'événements n.

La statistique de Shiryaev-Roberts spatio-temporelle est alors : $R_n = \sum_{k=1}^{n} \Lambda_{k,n}$. Une alarme est déclenchée quand R_n atteint le seuil d'alerte A fixé a priori à partir de l'ARL_0 que l'on souhaite garantir.

Pour localiser le cluster le plus probable, on recherche la valeur k^* telle que $k^* = \arg\max_k(\Lambda_{k,n})$ $(1 \leq k \leq n)$. Le cluster le plus probable est alors $z_{k^*,n}$, le cylindre centré sur le point (x_{k^*}, y_{k^*}), de rayon ρ et de hauteur $[t_{k^*}; t_n]$.

Cette méthode impose de recalculer les n $\hat{\Lambda}_{k,n}$ à chaque nouvelle observation. Afin d'éviter cela les auteurs démontrent qu'il est possible de calculer R_{n+1} en fonction des rapports de vraisemblance estimés précédemment. Pour plus de détail voir Assunção et Correa (2009) (30).

3.2.4 Tendance globale à l'agrégation

La méthode de suivi de la tendance globale à l'agrégation sur données agrégées (latticielles) a été proposée par Rogerson (1997) (31). Elle repose sur le suivi dans le temps d'un indicateur de cette tendance : la statistique de Tango. Cette statistique mesure la propension d'unités spatiales voisines à avoir ensemble un nombre élevé de cas. La proximité des unités spatiales est pondérée de façon à ce que l'importance donnée aux unités spatiales diminue de façon exponentielle avec leur éloignement.

On nomme r le vecteur de taille $M \times 1$ des proportions de cas observés dans chacune des M unités spatiales. Sous l'hypothèse nulle (H0), on suppose le risque de l'événement mesuré indépendant du lieu de vie. On construit donc le vecteur p des proportions attendues dans chacune des unités spatiales à partir de la répartition de la population exposée

dans les unités. Sous l'hypothèse alternative (H1), il existe au moins une région où le risque est augmenté par rapport au reste du territoire étudié.

Soit A la matrice de taille $M \times M$ dont les éléments $a_{i,j}$ constituent une mesure de proximité entre deux unités spatiales i et j. Tango calcule cette proximité en fonction de la distance $d_{i,j}$ ($i \neq j$) telle que : $a_{i,j} = \exp\left(\frac{d_{i,j}}{\tau}\right)$, avec τ un facteur d'échelle lié à la taille du cluster que l'on souhaite mesurer. L'augmentation de la valeur de τ permet de détecter des clusters de plus grande taille. En notation matricielle, la statistique de Tango C_G est alors :

$$C_G = (r-p)'A(r-p)$$

Son espérance $E(C_G)$ et sa variance $V(C_G)$ sont alors :

$$E(C_G) = \frac{1}{N}Tr(AV_p)$$

$$V(C_G) = \frac{2}{N^2}Tr(AV_p)^2$$

avec N le nombre total de cas observés et $V_p = \Delta(p) - pp'$ où $\Delta(p)$ est une matrice diagonale de taille $M \times M$ dont la diagonale contient les éléments de p.

Dans le cadre prospectif, on va s'intéresser à l'évolution de la tendance globale à l'agrégation au cours du temps. Pour cela, on va calculer un Z-score à partir de la statistique de Tango et de son espérance et de sa variance conditionnées sur les valeurs précédentes. Ce Z-score sera ensuite suivi au cours du temps dans le cadre d'une statistique de somme cumulée (CUSUM).

Soit r_{i-1} le vecteur de taille $M \times 1$ des proportions de cas observés dans chaque région après $i-1$ observations. On note $r_{i-1}(k)$ le vecteur des proportions de cas observés dans chaque région après i observations étant donné r_{i-1} et sachant que l'observation i est située dans l'unité spatiale k. L'espérance de la statistique de Tango $C_{G,i}$ après i observations, conditionnée par $C_{G,i-1}$ est : $E(C_{G,i}|C_{G,i-1}) = p'u$, avec u un vecteur de taille $M \times 1$ contenant pour chaque élément k ($1 \leq k \leq M$) : $u_k = (r_{i-1}(k) - p)'A(r_{i-1}(k) - p)$. La variance de $C_{G,i}$, conditionnée par $C_{G,i-1}$ est : $V(C_{G,i}|C_{G,i-1}) = p'diag(uu') - (p'u)^2$.

On peut donc construire le Z-score suivant :

$$Z_i = \frac{C_{G,i} - E(C_{G,i}|C_{G,i-1})}{\sqrt{V(C_{G,i}|C_{G,i-1})}}$$

Toutefois les Z_i n'ayant pas une distribution tout à fait normale, l'auteur propose de les réunir par groupe de n valeurs successives, puis d'utiliser la moyenne du groupe $\bar{Z}_{(n)}$ pour le calcul de la statistique de somme cumulative. Le nombre de Z-scores successifs dans le groupe doit être suffisant pour garantir la normalité de $\bar{Z}_{(n)}$, tout en restant le plus faible possible pour permettre la détection précoce, les tests réalisés par l'auteur lui permettent de recommender une valeur de $n = 4$.

La statistique de somme cumulative (CUSUM) S_i est alors définie par :

$$S_i = \text{max}(0, S_{i-1} + \bar{Z}_{(n),i} - k)$$

avec $S_0 = 0$ et k une valeur de référence, donnée en nombre d'écarts-types.

L'hypothèse nulle est rejetée si la statistique S_i dépasse le seuil critique h, fixé à l'avance, et le taux de faux positifs est contrôlé par le biais de l'average run length sous l'hypothèse nulle (ARL_0). Rogerson propose de fixer ce seuil à partir de tables pré-calculées permettant de faire la correspondance entre des valeurs de k et de l'ARL_0 et la valeur de h associée.

Il existe une version focalisée de la statistique de Tango permettant de détecter une tendance à l'agrégation autour d'une ou plusieurs unités spatiales pré-déterminées. Cette méthode peut également être suivie dans le temps par l'intermédiaire d'une statistique de somme cumulée, mais ne sera pas détaillée ici.

3.2.5 Méthodes de détection de cluster sur données agrégées

Plusieurs méthodes de détection de cluster ont été proposées. La plupart relèvent de la famille des statistiques de scan, qui consistent à balayer l'espace à la recherche de la localisation la plus probable pour un cluster éventuel. Une méthode complémentaire a également été produite par application de ce concept de balayage au cadre d'analyse des CUSUM.

L'hypothèse nulle (H0) est la distribution homogène des cas sur l'ensemble de l'espace et du temps. L'hypothèse alternative (H1) est l'existence d'un sous-ensemble z de $\{D, T\}$, le cluster, dans lequel la probabilité de survenue d'un événement serait augmenté, c'est-à-dire que la moyenne de la distribution serait supérieure à ce qui est attendu. Toutefois, ces hypothèses peuvent être interprêtées de deux façons.

La première, due à Kulldorff (32), interprête l'hypothèse alternative comme l'existence d'un cluster z tel que le risque de survenue de l'événement dans z est différent du risque de survenue dans z_c son complémentaire. Neill propose une interprétation légèrement différente (16,33). Lui interprète l'hypothèse alternative comme l'existence d'un cluster z tel que le risque de survenue de l'événement dans z est différent de ce qui était attendu sous H0. Il en résulte que là où Kulldorff compare le risque relatif (par rapport aux attendus) à l'intérieur du cluster au risque à l'extérieur, Neill va simplement comparer le risque relatif à l'intérieur du cluster à 1.

Les méthodes de détection de cluster permettant leur localisation se décomposent en 2 étapes :

1. Phase de détection
2. Phase d'inférence

3.2.5.1 Phase de détection La phase de détection consiste à parcourir l'espace et le temps de façon à couvrir tous les clusters possibles. Pour chaque cluster candidat, une statistique est calculée, qui correspond à un rapport de vraisemblance. Le mode de calcul de cette statistique dépend des hypothèses de chaque modèle proposé. A l'issue de cette phase de détection, on cherche à identifier le cluster candidat pour lequel la statistique est maximale, c'est le cluster le plus probable. Dès lors, on peut distinguer deux approches : celle de la statistique de scan (ou de balayage), et celle de la CUSUM de balayage.

Dans les deux cas on va commencer par un processus de balayage des deux dimensions d'espace à la recherche de toutes les combinaisons d'unités spatiales contiguës possibles compte tenu des contraintes que l'on se fixe a priori. En effet, le nombre de ces combinaisons possible est très élevé, et on est généralement contraint de limiter les formes de cluster acceptables pour contrôler le temps de calcul associé.

La forme de fenêtre de balayage la plus couramment adoptée est la forme circulaire, essentiellement en raison de la rapidité de l'algorithme de recherche associé (34). Soit z une fenêtre circulaire de taille variable, elle prend comme centre un site d_i de D et se déplace sur l'ensemble de l'espace à 2 dimensions pour constituer une collection Z de tous les clusters potentiels. La fenêtre étant de taille variable, pour chaque centre d_i tous les rayons seront testés de façon à inclure progressivement tous les autres points de D. En pratique on limite généralement la taille de la fenêtre de façon à ce qu'elle ne contienne plus de K unités spatiales, ou 50 % des événements. Là encore cette limitation peut être destinée à réduire le nombre de clusters candidats, et donc de calculs à effectuer, mais elle prend aussi tout son sens lorsque l'on cherche à comparer le risque à l'intérieur du cluster au risque dans le reste du territoire. En effet, dans ce cas, détecter un cluster de sur-risque z regroupant plus de 50 % des cas revient à détecter un cluster de sous-risque z_C complémentaire à z.

Toutefois, d'autres formes ont été envisagées. Ainsi, Takahashi et al. (35) ont proposé une fenêtre de scan dont la partie spatiale est de forme variable : elle prend aussi comme centre un site d_i de D et se déplace sur l'ensemble du graphe G de façon à constituer une collection Z de clusters potentiels. Mais pour chaque site d_i parcouru, la fenêtre inclut l'ensemble des cercles concentriques de ses K plus proches voisins (comme la statistique de scan circulaire), mais aussi tous les ensembles d'unités spatiales connectées entre elles (suivant le graphe de voisinage G) et comprenant d_i parmi ses K plus proches voisins. Une autre alternative est la proposition de Neill (33). Dans ce cas, On agrège les unités spatiales d dans une grille G_N bidimensionnelle uniforme de taille $N \times N$. Soit z une fenêtre rectangulaire de taille variable, elle parcourt l'ensemble des régions rectangulaires de la grille G_N. Ces deux méthodes nécessitent davantage de temps de calcul qu'une fenêtre circulaire, mais Neill propose un algorithme de recherche accéléré pour sa méthode

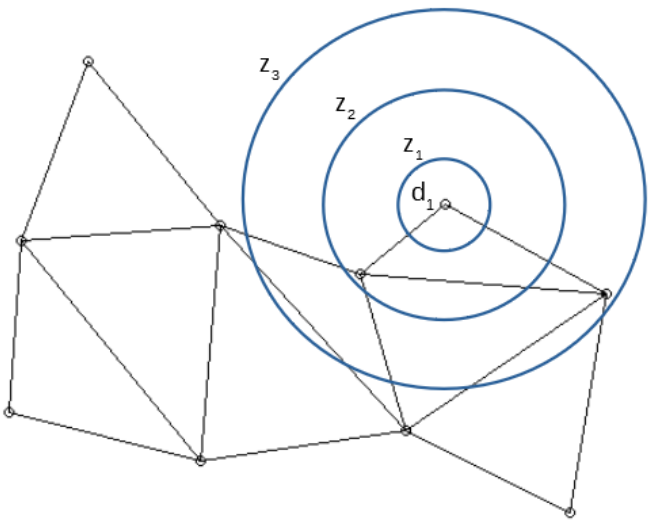

FIG. 3 : Illustration du processus de scan spatial pour la région Grand-Est. Pour chaque sommet d du graphe de voisinage une fenêtre circulaire de taille croissante couvre progressivement les points les plus proches afin de constituer une collection de clusters candidats.

(36). Enfin, Sonesson (22) a également proposé une fenêtre “de voisinage”, c’est-à-dire qu’elle regroupe l’unité spatiale sur laquelle elle est centrée, ainsi que toutes les unités adjacentes.

Les clusters que l’on cherche à identifier n’ont pas qu’une dimension spatiale, mais aussi temporelle. Se situant dans le cas prospectif, on s’intéresse exclusivement aux clusters encore actifs, c’est-à-dire qui comprennent le temps t où l’analyse est réalisée. On cherche donc à identifier le temps τ où le cluster a le plus probablement débuté, sa date de fin étant fixée à t. Soit $LR(z,\tau)$ la statistique de rapport de vraisemblance associée au cluster candidat z pour lequel la date de début serait τ, et $LR(z)$ son maximum pour un z donné, on a donc :

$$LR(z) = \max_{\tau} LR(z,\tau)$$

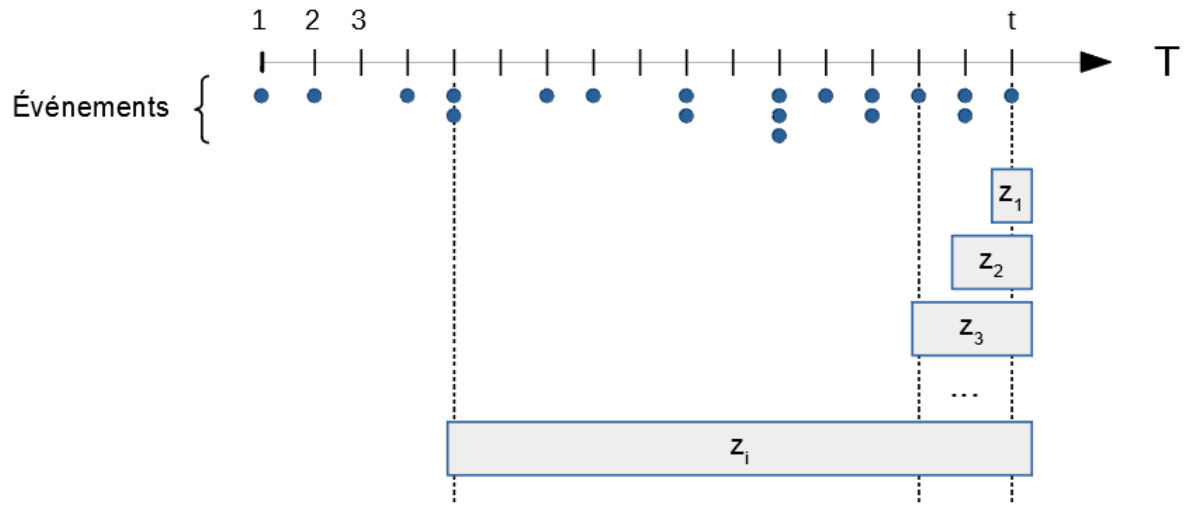

FIG. 4 : Illustration du processus de scan temporel prospectif en temps discret. Pour une analyse réalisée au temps t, toutes les fenêtres temporelles se terminant au temps t sont parcourues de façon à calculer le rapport de vraisemblance correspondant.

C’est sur la méthode utilisée pour rechercher $LR(z)$ que l’approche par statistique de scan diffère de celle de la CUSUM de balayage. En effet, avec la méthode de statistique de scan on calcule toutes les valeurs possibles de $LR(z,\tau)$, puis on en conserve le maximum. Cela revient à balayer les 3 dimensions d’espace-temps avec une fenêtre de forme cylindrique, dont la base serait la fenêtre circulaire qui parcourt les deux dimensions d’espace, tandis que la hauteur parcourt le temps. La méthode de CUSUM de balayage procède différemment. Sonesson a montré, à partir de résultats antérieurs (22), que si $LR_t(z)$ est la valeur de $LR(z)$ quand on réalise l’analyse au temps t, alors :

$$LR_t(z) = \max_{\tau} LR_t(z, \tau) = \max(0, LR_{t-1}(z) + \ln(LR_t(z, t)))$$

Autrement dit, le processus de scan temporel prospectif est équivalent à une procédure de CUSUM de rapports de vraisemblance où, à chaque temps t, on ajoute à la valeur précédente de la statistique de test le logarithme du rapport de vraisemblance associé aux données observées au temps t. Avec la méthode de CUSUM de balayage, on ne calcule donc qu'une seule statistique de test par cluster candidat, via la mise à jour de sa valeur au temps $t-1$.

A l'issue de la phase de détection on conserve, parmi l'ensemble Z des clusters candidats, celui qui maximise LR(z). La statistique de test globale est donc :

$$LR = \max_{z \in Z} LR(z)$$

C'est cette statistique qui sera testée lors de la phase d'inférence, on réalise donc 1 seul test. Le cluster correspondant est considéré comme le plus probable (most likely cluster – MLC).

Le calcul de la statistique de test associée à chaque cluster candidat nécessite de faire une hypothèse quant à la distribution des données sous H0 et sous H1. Plusieurs modèles ont été proposés dans ce contexte :

- la statistique de scan prospective de Kulldorff classique, ou conditionnelle pour distribution de Poisson (32)
- la statistique de scan par permutations (37)
- la statistique de scan non conditionnelle pour distribution de Poisson (33)
- la statistique de scan non conditionnelle pour distribution binomiale négative (38)
- la statistique de scan non conditionnelle pour distribution de Poisson zéro-inflatée (39)
- la CUSUM circulaire par balayage ou C-CUSUM (22)

3.2.5.1.1 Méthode conditionnelle de Poisson On se place dans le cas de la méthode conditionnelle pour modèle de Poisson proposée par Kulldorff (32). Soit une variable aléatoire X(d, t) correspondant au nombre de personnes concernées par l'événement étudié dans l'unité d au temps t. Pour $\forall(d,t) \in \{D,T\}$, $X(d,t) \sim P(p\mu(d,t))$ où $\mu(d,t)$ correspond à une mesure d'intensité caractérisant la population sous-jacente à l'unité d au temps t. On rappelle que son hypothèse alternative est l'existence d'un cluster z tel que le risque de survenue de l'événement à l'intérieur du cluster (p_z) est différent de celui observé dans le reste du territoire (p_{z_c}). Sous H0, ce risque est présumé constant (p_0). Ces paramètres sont estimés par maximum de vraisemblance lors du calcul de la statistique. Si L_{H0} et L_{H1} sont les vraisemblances sous l'hypothèse nulle et sous l'hypothèse alternative respectivement, on a :

$$LR = \max_{z \in Z} \frac{\max_{p_z ; p_{z_c}} L_{H1}}{\max_{p_0} L_{H0}}$$

Si $x(z)$ est le nombre de cas observés dans z, et $x(D)$ le nombre total de cas observés dans le territoire étudié (D), alors la statistique de test globale est :

$$LR = \max_{z \in Z} \frac{p_z^{x(z)} p_{z_c}^{x(z_c)}}{p_0^{x(D)}} = \max_{z \in Z} \frac{\left(\frac{x(z)}{\mu(z)}\right)^{x(z)} \left(\frac{x(z_c)}{\mu(z_c)}\right)^{x(z_c)}}{\left(\frac{x(D)}{\mu(D)}\right)^{x(D)}}$$

Le paramètre $\mu(d,t)$ correspond généralement à la population des unités spatiales, mais il peut aussi correspondre à un nombre de cas attendu. On remarque que dans ce cas, p_z ne représente plus le risque de survenue de l'événement, mais un facteur de proportionnalité reliant le nombre de cas attendus au nombre observé, la statistique de scan évalue alors si ce facteur varie significativement dans un cluster candidat par rapport au reste de l'espace-temps étudié. Cette méthode a l'intérêt d'être relativement résistante à une mauvaise estimation du nombre d'événements attendu, à partir du moment où le biais est systématique et ne dépend pas de l'espace ou du temps.

3.2.5.1.2 Méthode par permutations Cette première méthode nécessite d'avoir des informations sur la population sous-jacente ou des données historiques sur le nombre de cas afin d'estimer les attendus. Mais parfois ces informations ne sont pas disponible, par exemple en cas de méconnaissance de la part d'une population réellement exposée à un risque, ou de la part de la population privilégiant un hôpital donné en cas de maladie en zone urbaine (un même individu ayant le choix entre plusieurs établissements proches). Pour répondre à cette problématique, Kulldorff et al. (2005) (37) ont proposé une variante.

Un nombre de cas attendu $\mu(d,t)$ est calculé exclusivement à partir des cas observés pendant la période étudiée (voir section 3.2.6 sur les méthodes de calcul des attendus), puis on reprend les mêmes hypothèses que précédemment. Désormais, le nombre de cas observés dans un cylindre z ($x(z)$) est supposé suivre une loi hypergéométrique de moyenne $\mu(z)$. Toutefois, lorsqu'à la fois le nombre de cas observés au cours de la même période que z, et le nombre de cas observés dans la même fenêtre spatiale que z sont petits par rapport au nombre total de cas observés $x(D)$, il est possible d'approcher la distribution de $x(z)$ par une distribution de Poisson de moyenne $\mu(z)$. On retrouve alors la statistique précédente, mais sachant que le nombre d'attendus est calculé à partir des observés, on a $\mu(D) = x(D)$ et le dénominateur devient égal à un, ce qui donne :

$$LR = \max_{z \in Z} p_z^{x(z)} p_{z_c}^{x(z_c)} = \max_{z \in Z} \left(\frac{x(z)}{\mu(z)}\right)^{x(z)} \left(\frac{x(z_c)}{\mu(z_c)}\right)^{x(z_c)}$$

Il faut noter que ces statistiques ont été formulées de façon à évaluer la vraisemblance de la différence constatée entre le risque dans z et le risque sur le reste l'espace étudié, sachant que le nombre total d'événements est connu et égal à $x(D)$. Autrement dit, la statistique est calculée comme si on savait su, avant même la réalisation de l'analyse, combien de cas seraient obtenus au total, c'est pourquoi ces méthodes sont dites conditionnées (sur $x(D)$). L'utilisation d'une telle approche, dans le cadre prospectif où par définition le nombre de cas qui devra être analysé est inconnu, a été critiquée par certains auteurs (40), qui dans ce cadre recommendent plutôt une approche non conditionnelle. Le nombre total de cas observé est alors considéré comme une variable aléatoire, donc inconnue a priori.

3.2.5.1.3 Méthode non conditionnelle de Poisson La méthode proposée par Neill entre dans ce cadre. Par ailleurs, elle correspond aussi à une interprétation différente de l'hypothèse alternative, dans laquelle on compare directement le risque de survenue de l'événement dans le cluster (p_z) à 1. Cela signifie que les nombres de cas attendus sous l'hypothèse nulle ($\mu(d,t)$) sont supposés correspondre réellement aux moyennes des distributions de Poisson dans chaque unité spatiale, et p_z ne représente plus le risque de survenue de l'événement dans z, mais son risque relatif dans z par rapport au niveau de base. Ainsi, là où Kulldorff estime à la fois p_z, p_{z_c} et p_0 par maximum de vraisemblance, Neill n'estime qu'un seul paramètre : p_z, soit :

$$LR = \max_{z \in Z} \frac{\max_{p_z > 1} L_{H1}}{L_{H0}}$$

Ce qui, pour une distribution de Poisson, donne :

$$LR = \max_{z \in Z} \max_{p_z > 1} e^{(1-p_z)\mu(z)} p_z^{x(z)} = \max_{z \in Z} \left(\frac{x(z)}{\mu(z)} \right)^{x(z)} e^{\mu(z)-x(z)}$$

Et la valeur de p_z qui maximise le numérateur est : $p_z = \max\left(1, \frac{x(z)}{\mu(z)}\right)$.

La méthode de Neill étant fondée sur la seule évaluation du risque relatif dans le cluster candidat, elle est moins résistante que celle de Kulldorff à une mauvaise estimation systématique du risque de base.

3.2.5.1.4 Méthode non conditionnelle binomiale négative Dans le contexte de la surveillance épidémiologique, la variance empirique des nombres de cas observés dans le temps peut parfois être supérieure à leur moyenne empirique, ce qui correspond à une situation de surdispersion par rapport à une loi de Poisson où moyenne et variance sont égales. Dans ce cas Tango et al. (38) suggèrent de préférer une loi binomiale négative à la loi de Poisson, qui est caractérisée par sa moyenne μ et un paramètre ϕ qui contrôle la surdispersion.

Pour $\forall(d,t) \in \{D,T\}$, $X(d,t) \sim NB(\mu(d,t),\phi(d,t))$ où $\mu(d,t)$ correspond à l'espérance du nombre de cas dans l'unité d au temps t. La surdispersion temporelle est alors $w(d,t) = 1 + \mu(d,t)/\phi(d,t)$. Sous l'hypothèse alternative (H1), il existe un cluster z tel que $\forall(d,t) \in z$, $X(d,t) \sim NB(p_z\mu(d,t),\phi(d,t))$, avec $p_z > 1$ le risque relatif inconnu dans le cluster z.

Pour le calcul de la statistique de test, les auteurs distinguent deux cas de figure quant au mode d'apparition du cluster :

- le modèle de point chaud, où le risque relatif est constant sur l'ensemble de la durée du cluster
- le modèle épidémique, où le risque relatif augmente avec le temps dans le cluster

Si l'on se place dans le cadre du modèle épidémique, pour un cluster se terminant au temps t_p et d'une durée u, on peut exprimer p_z comme une fonction du temps qui vaut 1 au temps $t = t_p - u$ (le dernier temps avant le début du cluster), et qui augmente de façon monotone ensuite. On s'intéresse à la pente initiale de cette fonction, qui dépend d'un facteur β_z. Tester l'hypothèse alternative revient donc à tester si $\beta_z > 0$.

Mais la construction du rapport de vraisemblance nécessite la connaissance de la forme fonctionnelle de cette fonction, qui peut prendre des formes différentes selon la situation, ainsi qu'un estimateur de maximum de vraisemblance pour β_z. Les auteurs proposent donc une statistique de test qui ne dépend ni de l'un ni de l'autre et qui est asymptotiquement équivalente au test de rapport de vraisemblance :

$$LR = \max_{z\in Z, 1\leq u\leq T} \frac{\sum_{(d,t)\in z}(x(d,t)-\mu(d,t))(t-t_p+u)/w(d,t)}{\sqrt{\sum_{(d,t)\in z}\mu(d,t)(t-t_p+u)^2/w(d,t)}}$$

Si l'on se place dans le cadre du modèle de point chaud, pour un cluster z on a un risque relatif $p_z > 0$ constant sur l'ensemble du cluster. La statistique de test devient alors :

$$LR = \max_{z\in Z, 1\leq u\leq T} \frac{\sum_{(d,t)\in z}(x(d,t)-\mu(d,t))/w(d,t)}{\sqrt{\sum_{(d,t)\in z}\mu(d,t))/w(d,t)}}$$

Ces statistiques peuvent s'adapter au modèle de Poisson, il suffit alors d'attribuer la valeur 1 à $w(d,t)$ pour toutes les unités spatiales à tous les temps. Dans le cas du modèle de point chaud on a alors :

$$LR = \max_{z\in Z} \frac{x(z)-\mu(z)}{\sqrt{\mu(z)}}$$

3.2.5.1.5 Méthode non conditionnelle pour distribution ZIP Dans d'autres situations, on peut être confronté à une forte sur-représentation de la valeur "0" dans les données, ce qui ne permet plus d'assimiler leur distribution à la distribution de Poisson. Dans ce cas, Allévius et Höhle (39) proposent d'utiliser plutôt une distribution de Poisson zéro-inflatée. Dans le domaine de la surveillance épidémiologique, plusieurs raisons peuvent conduire à l'obtention de zéro événement pour une unité spatiale et un temps donnés. Une première est qu'aucun des patients malades n'a consulté de médecin, ce cas correspond au "zéro d'échantillonnage". Une autre cause serait l'absence de patient atteint et de faux positif, ou l'incapacité continue du système de surveillance à détecter les cas, ce qui correspond au "zéro structurel". La distribution de Poisson zéro-inflatée (ZIP) permet de prendre en compte les zéros structurels.

On considère que X suit une loi ZIP, qui est caractérisée par 2 paramètres : q la probabilité d'obtenir un zéro structurel, et μ l'espérance de la distribution lorsque la valeur de X n'est pas un zéro structurel. Sous l'hypothèse nulle, pour $\forall(d,t) \in \{D,T\}, X(d,t) \sim ZIP(q(d,t),\mu(d,t))$ où $\mu(d,t)$ correspond à l'espérance du nombre de cas dans l'unité d au temps t en l'absence de zéro structurel. Sous l'hypothèse alternative, si z est un cluster alors pour $\forall(d,t) \in z, X(d,t) \sim ZIP(q(d,t), p_z\mu(d,t))$, avec p_z le risque relatif dans le cluster z. Tester l'hypothèse alternative revient donc à tester si $p_z > 1$.

Soit $\delta(d,t)$ un indicateur de zéro structurel qui prend la valeur 1 quand le compte dans l'unité d au temps t est un zéro structurel, et 0 sinon. Habituellement $\delta(d,t)$ est inconnu et, comme p_z, il ne peut être obtenu par une solution analytique. Ces deux paramètres sont donc estimés ensemble par application d'un algorithme de type *expectation-maximization* (EM), et ce pour chaque cluster candidat, à partir des paramètres $\widehat{q}$ et $\widehat{\mu}$ de la distribution ZIP estimés par régression sur des données historiques.

Une fois tous les paramètres estimés, il est possible de calculer la statistique de scan à partir des vraisemblances associées à chaque cluster candidat :

$$\begin{aligned} LR &= \max_{z\in Z} \log\left(\frac{L(\hat{p}_z)}{L(1)}\right) \\ &= \max_{z\in Z} \log\left(\frac{\prod_{d,t} Pr(X(d,t)=x(d,t)|\hat{q}(d,t),\hat{p}_z\hat{\mu}(d,t))}{\prod_{d,t} Pr(X(d,t)=x(d,t)|\hat{q}(d,t),\hat{\mu}(d,t))}\right) \end{aligned}$$

Sachant que la probabilité est donnée par :

$$Pr(X(d,t)=x(d,t)) = \begin{cases} \frac{\exp(-p_z\mu(d,t))(p_z\mu(d,t))^{x(d,t)}}{x(d,t)!} & \text{si } \delta(d,t)=0 \\ 1, & \text{si } \delta(d,t)=1 \end{cases}$$

3.2.5.1.6 Méthode de CUSUM de balayage Nous avons vu qu'avec les approches de Kulldorff on ne fait pas d'hypothèse concernant la valeur de p_z, p_{z_c} et p_0, ceux-ci sont

estimés dans le rapport de vraisemblance. Avec les différentes approches non conditionnelles, on fait une hypothèse sur la valeur du risque sous H0, mais pas sous l'hypothèse alternative. L'approche par C-CUSUM, ou CUSUM de balayage, proposée par sonesson (22) complète l'ensemble en imposant de faire une hypothèse sur le risque à la fois sous H0 et sous H1.

Soient $\lambda_0(d,t) = p_0 x(d,t)$ le nombre de cas attendus dans l'unité spatiale d au temps t sous l'hypothèse nulle H0, exprimé en fonction de p_0 le taux d'événements commun à toutes les unités sous H0, et $\lambda_1(d,t) = p_z x(d,t)$ le nombre de cas attendus dans l'unité spatiale d au temps t sous l'hypothèse alternative H1 où cette unité ferait partie du cluster z, avec p_z le taux d'événements correspondant. On appellera $\lambda_0^*(d,t)$ et $\lambda_1^*(d,t)$ les hypothèses fixées pour ces valeurs et utilisées dans le modèle.

Soit $LR_t(z, \tau, \lambda_1^*, \lambda_0^*)$ le rapport de vraisemblance partiel pour un changement dans un processus de Poisson d'un niveau spécifié λ_0^* à un nouveau niveau spécifié λ_1^*, ayant lieu au temps $T = \tau$ dans une zone spécifiée z, l'analyse étant réalisée au temps t.

On définit la CUSUM pour un cluster candidat z au temps t par :

$$LR_t(z) = \max_{\tau} LR_t(z,\tau) = \max(0, LR_{t-1}(z) + \ln(LR_t(z,t)))$$

Ce qui donne :

$$LR_t(z) = \max\left(0, LR_{t-1}(z) + \sum_{d \in z} x(d,t) \ln \frac{\lambda_1^*(d,t)}{\lambda_0^*(d,t)} - \sum_{d \in z} (\lambda_1^*(d,t) - \lambda_0^*(d,t))\right)$$

Et la statistique de test globale au temps t est donc :

$$LR_t = \max_{z \in Z} LR_t(z)$$

3.2.5.2 Phase d'inférence La phase d'inférence statistique consiste à évaluer si la statistique obtenue est significativement différente de ce que l'on pourrait obtenir sous l'hypothèse nulle d'absence de cluster. Pour cela on doit comparer la valeur obtenue à la distribution sous H0.

Toutefois la loi de probabilité de LR sous H0 ne peut être obtenue sous forme analytique en raison de la dépendance causée par les chevauchements multiples des fenêtres, on utilise donc la méthode de Monte Carlo pour obtenir une approximation de sa distribution. On réalise R simulations sous H0 avec phase de détection du MLC afin d'obtenir l'ensemble Λ des LR_i ($1 \leq i \leq R$) simulés. La probabilité critique P est estimée par le rapport entre le nombre de LR_i simulés supérieurs ou égaux au LR_0 du jeu de données

étudié, et $R+1$. H0 est rejetée si P est inférieur au risque de première espèce α fixé (par exemple 5 %).

Ces simulations sont réalisées conformément à l'hypothèse de distribution sous H0 du modèle considéré, mais on peut distinguer deux approches quant au calcul des cas observés simulés selon que le modèle est conditionnel ou non. Dans le cas conditionnel, le nombre total de cas à simuler est connu : il s'agit du nombre total de cas observés en réalité, les événements sont donc répartis dans chaque unité spatiale de façon à maintenir un nombre total identique dans toutes les simulations et le jeu de données réel. L'approche non conditionnelle, elle, considère le total des observés comme une variable aléatoire, donc inconnue a priori. Les cas sont alors simulés indépendamment dans chaque unité spatiale, sans contrainte sur le nombre total devant être atteint.

Par ailleurs, la méthode se scan par permutations se distingue de la méthode classique de Kulldorff à cette étape. En effet, là où la méthode conditionnelle pour distribution de Poisson répartit les cas observés simulés entre les unités spatiales en fonction de leurs populations respectives, la méthode par permutations ne peut procéder de la même façon en l'absence d'informations sur la population sous-jacente. C'est pourquoi les simulations conditionnées sur le nombre total de cas observés sont réalisées par permutations des dates et des localisations des nombres de cas observés en réalité, en maintenant les sommes marginales par unité spatiale et par date identiques.

Une alternative à la méthode de Monte Carlo a été proposée par Neill (33). En effet, celui-ci remarque que, la méthode de détection ne permettant pas de prendre en compte l'incertitude sur les estimations des attendus, lorsque la variance de ces estimations augmente le taux de faux positifs n'est plus contrôlé et peut devenir très important. Il propose alors deux solutions. La première maintient le principe de la méthode de Monte Carlo, mais avec un seuil de risque de première espèce plus faible, ce qui a deux inconvénients : le risque α nominal ne correspond plus au risque α empirique, et cela nécessite une augmentation considérable du nombre de simulations à effectuer, et donc un temps de calcul beaucoup plus important. Sa seconde solution consiste à remplacer les simulations de Monte Carlo par l'utilisation de la distribution empirique des statistiques de scan observées sur des données historiques, puis de calculer le seuil critique de la même façon qu'avec les données simulées. Les inconvénients de cette méthode sont qu'elle nécessite une grande quantité de données historiques et qu'elle suppose une absence de cluster au sein de ces données.

Dans le cas prospectif les analyses de détection de cluster ont vocation à être répétées régulièrement au fil du temps, il faut alors prendre en compte ces tests multiples dans l'évaluation de la significativité du test. Kulldorff (32) a proposé de corriger la p-valeur obtenue par simulation de Monte Carlo pour ces tests multiples en élargissant la recherche de cluster sur ces simulations aux clusters étudiés lors des analyses précédentes. La phase de détection inclut alors tous les clusters candidats des analyses antérieures, c'est-à-dire tous ceux qui se terminent à une date comprise entre T_1 celle de la première analyse et T_n la dernière date disponible. Cette méthode étant particulièrement conservatrice, il propose de la moduler en fixant une fenêtre temporelle de taille w sur laquelle on souhaite garantir le risque de première espèce, on inclut donc tous les clusters se terminant entre T_{n-w} et

T_n. Suite aux critiques de Correa (41) concernant le fait que cette méthode implique que le risque α nominal ne corresponde plus au risque α empirique pour une analyse donnée, Kulldorff précise que sa méthode vise à préserver le risque de première espèce nominal sur une période donnée, et que par conséquent elle est effectivement équivalente à une simple réduction du seuil de risque α pour chaque analyse prise individuellement (42).

Une autre approche est celle de la CUSUM, qui consiste à fixer le risque de faux positifs par le biais de l' *average run length* sous l'hypothèse nulle (ARL_0), c'est-à-dire le nombre moyen d'analyses successives nécessaire pour obtenir un premier faux positif en l'absence de cluster (22). On réalise alors R simulations sous l'hypothèse nulle, puis des analyses sont réalisées successivement à plusieurs temps t pour chacun de ces jeux de données. Plusieurs seuils de score sont alors testés afin de calculer l' ARL_0 correspondant jusqu'à obtention de la valeur désirée. Le même seuil est ensuite conservé pour toutes les analyses successives.

3.2.6 Calcul du nombre de cas attendus

Les méthodes de détection de cluster se basent sur la comparaison du nombre d'événements observés au nombre d'événements qui serait attendu en l'absence de phénomène d'agrégation des cas. Les auteurs proposent plusieurs méthodes d'estimation des nombres de cas attendus (16,33). La plupart sont non biaisées mais certaines, se basant sur le maximum observé parmi les valeurs antérieures suivant une fenêtre glissante, sont biaisées positivement. Toutefois les simulations réalisées dans ces articles les conduisant à ne pas recommander l'utilisation d'un maximum, nous ne traiterons ici que les méthodes non biaisées.

Les différentes méthodes proposées sont :

- les modèles de régression
- la moyenne glissante suivant une fenêtre temporelle fixée
- la moyenne glissante à lissage exponentiel suivant une fenêtre temporelle fixée
- la méthode multiplicative de Holt-Winters, variante de la précédente
- la méthode dite du "jour actuel", proposée par Kulldorff (2005) (37)

Dans la plupart des cas, on fera l'hypothèse d'une distribution de Poisson sous l'hypothèse nulle, Allévius (39) propose alors simplement d'estimer le nombre de cas attendus par le biais d'une régression de Poisson. Cette méthode a l'avantage de permettre de prendre en compte divers facteurs d'ajustement relatifs aux unités spatiales étudiées (milieu urbain ou rural par exemple) ou aux populations sous jacentes. Lorsque l'on souhaite pouvoir prendre en compte une éventuelle surdispersion des données, donc une situation la variance empirique des nombres de cas observés dans le temps est supérieure à leur moyenne empirique, Tango et al. (38) suggèrent de préférer une loi binomiale négative à la loi de Poisson, qui est caractérisée par sa moyenne μ et un paramètre ϕ qui contrôle la surdispersion. Dans d'autres situations, on peut être confronté à une forte sur-représentation de la valeur "0" dans les données, ce qui ne permet plus d'assimiler leur distribution à la distribution de Poisson. Dans ce cas, Allévius et Höhle (39) proposent d'utiliser plutôt une distribution de Poisson zéro-inflatée. S'il existe des données antérieures sur une durée

suffisamment longue et dénuée d'épidémie, les paramètres de cette distribution peuvent être estimés par le biais d'un modèle de régression de Poisson zéro-inflatée.

La moyenne glissante sur n jours consiste simplement à estimer le nombre de cas attendu au jour t en calculant la moyenne sur les n jours précédents : $\mu(d,t) = \frac{1}{n}\sum_{t-n\leq i\leq t-1} x(d,t)$. Il y a deux méthodes permettant de prendre en compte les variations quotidiennes : la stratification et l'ajustement (16). Dans le cas de la stratification on réalise le même calcul non pas à partir des n derniers jours, mais à partir des valeurs observées le même jour de la semaine, au cours des n dernières semaines. L'estimation est donc réalisée à partir de 7 séries temporelles en parallèle pour chaque jour de la semaine. Pour la méthode d'ajustement on considère que le nombre de cas observé un jour donné est le produit d'un nombre de cas de base par une constante dépendant du jour de la semaine. On calcule sur données antérieures la proportion β_i ($1 \leq i \leq 7$) de cas observés chaque jour de la semaine. Puis on divise les comptes observés au cours des n jours précédents par $7\beta_i$, le facteur correspondant au jour i. On réalise alors un calcul de moyenne glissante sur les n derniers jours, et on obtient le nombre attendu en multipliant le résultat par $7\beta_i$.

Une variante de cette moyenne glissante est proposée par Tango (38) afin de s'adapter à une distribution binomiale négative du nombre de cas. Pour cela on définit une période de B unités de temps et pour chaque temps t on calcule, à partir de la moyenne empirique $\bar{x}(d,t)$ et la variance empirique $s^2(d,t)$ estimés dans l'unité d sur l'intervalle $[t-B;t]$, les paramètres du modèle :

$$\mu(d,t) = \hat{\mu}^{(X)}(d,t) = \bar{x}(d,t)$$

$$\phi(d,t) = \hat{\phi}^{(X)}(d,t) = \begin{cases} \frac{\bar{x}^2(d,t)}{s^2(d,t)-\bar{x}(d,t)}, & \text{si } s^2(d,t) > \bar{x}(d,t) \\ \infty, & \text{sinon} \end{cases}$$

où $\phi(d,t)$ correspond au paramètre de surdispersion du modèle évoqué précédemment.

La moyenne glissante à lissage exponentiel sur n jours utilise une formulation récursive de façon à pondérer les données antérieures pour leur donner une importance décroissant de façon exponentielle avec l'ancienneté. Soit $\mu_0(d,t) = x(d,t-n)$ la valeur initiale de la formule récursive, on calcule la valeur attendue au temps t à partir des n valeurs antérieures par : $\mu(d,t) = \mu_t(d,t)$ avec $\mu_i(d,t) = \alpha x(d,t-n+i) + (1-\alpha)\mu_{i-1}(d,t)$ pour $1 \leq i \leq n-1$ et α un facteur de lissage. Cette méthode permet de prendre en compte les variations quotidiennes de la même façon que pour la précédente.

La méthode multiplicative de Holt-Winters est une variante de la moyenne glissante à lissage exponentiel dont Burkom et al. (43) ont montré l'intérêt pour la surveillance en contexte biologique ou de santé. Elle permet de prendre en compte les variations quotidiennes liées au jour de la semaine ainsi que les tendances linéaires de plus long terme. Elle est calculée également de façon récursive en itérant les 3 équations suivantes pour

la valeur lissée S_i, la composante de tendance T_i et la composante cyclique I_i ($t-n \leq i \leq t-1$) :

$$S_i = \alpha \frac{x(d,i)}{I_{i-7}} + (1-\alpha)(S_{i-1} + T_{i-1})$$

$$T_i = \beta(S_i - S_{i-1}) + (1-\beta)T_{i-1}$$

$$I_i = \gamma \frac{x(d,i)}{S_i} + (1-\gamma)I_{i-7}$$

On peut alors calculer le nombre de cas attendu $\mu(d,t) = (S_t + T_t)I_{t-7}$.

La méthode du "jour actuel" a été proposée par Kulldorff pour l'estimation des attendus dans la statistique de scan par permutations (37). Dans ce cas on suppose les cas distribués indépendamment dans le temps et dans l'espace, donc le nombre de cas attendus au temps t dans l'unité d est égal au produit du nombre de cas au temps t par le nombre de cas dans l'unité d, divisé par le nombre total de cas. Le nombre de cas attendus est donc calculé à partir du nombre de cas observés le même jour selon la formule :

$$\mu(d,t) = \frac{\sum_d x(d,t) \sum_t x(d,t)}{\sum_{(d,t)} x(d,t)}$$

Les auteurs ne mettent pas en évidence de méthode qui soit clairement supérieure aux autres toutefois ils recommendent d'utiliser une fenêtre temporelle d'au moins 28 jours pour les moyennes glissantes. Par ailleurs, ils constatent que les méthodes pour lesquelles l'estimation est associée à une plus forte variance conduisent à une dégradation du délai de détection lorsque l'on n'admet que de faibles taux de faux positifs (16). C'est le cas notamment des méthodes prenant en compte les variations associées au jour de la semaine. Il faut dès lors utiliser une plus grande fenêtre temporelle pour mieux contrôler la variance (par exemple 28 semaines au lieu de 28 jours pour une moyenne glissante stratifiée sur le jour de la semaine).

3.2.7 La statistique de scan bayésienne

Une version bayésienne des statistiques de scan évoquées précédemment a été proposée par Neill (44). Le processus de balayage est réalisé de la même façon, et on reprend ici les mêmes hypothèses et notations, avec une hypothèse de distribution de Poisson pour les nombres de cas observés, et p_z, p_{z_c} et p_0 les risques de survenue de l'événement dans le cluster candidat z, en-dehors de z et sous l'hypothèse nulle respectivement.

On considère que les taux de survenue suivent une distribution Gamma de telle sorte que sous l'hypothèse nulle $p_0 \sim Ga(\alpha_0, \beta_0)$, et que sous l'hypothèse alternative H1(z)

$\forall(d,t) \in z$ on a $p_z \sim Ga(\alpha_z, \beta_z)$, et $\forall(d,t) \notin z$ on a $p_{z_c} \sim Ga(\alpha_{z_c}, \beta_{z_c})$. Les paramètres $\alpha_.$ et $\beta_.$ des distributions Gamma sont fixés a priori par l'utilisateur, nous y reviendrons.

Dans ce contexte et par application du théorême de Bayes, les probabilités a posteriori des hypothèses étant donné un jeu de données D, sont données par $Pr(H0|D) = \frac{Pr(D|H0)Pr(H0)}{Pr(D)}$ pour l'hypothèse nulle, et par $Pr(H1(z)|D) = \frac{Pr(D|H1(z))Pr(H1(z))}{Pr(D)}$ pour l'hypothèse alternative. Les valeurs a priori $Pr(H0)$ et $Pr(H1(z))$ des probabilités de survenue des hypothèses sont fixées par l'utilisateur, et $Pr(D) = Pr(D|H0)Pr(H0) + \sum_{z \in Z} Pr(D|H1(z))Pr(H1(z))$ avec Z l'ensemble de tous les clusters candidats z.

A partir de ces hypothèse, Neill parvient à donner une formule de calcul des probabilités d'observer les données sachant l'hypothèse nulle :

$$Pr(D|H0) \propto \frac{{\beta_0}^{\alpha_0}\Gamma(\alpha_0 + x(D,T))}{(\beta_0 + \mu(D,T))^{\alpha_0 + x(D,T)}\Gamma(\alpha_0)}$$

avec $\Gamma(\cdot)$ la fonction gamma, et la probabilité équivalente sous l'hypothèse alternative :

$$Pr(D|H1(z)) \propto \frac{{\beta_z}^{\alpha_z}\Gamma(\alpha_z + x(z))}{(\beta_z + \mu(z))^{\alpha_z + x(z)}\Gamma(\alpha_z)} \times \frac{{\beta_{z_c}}^{\alpha_{z_c}}\Gamma(\alpha_{z_c} + x(z_c))}{(\beta_{z_c} + \mu(z_c))^{\alpha_{z_c} + x(z_c)}\Gamma(\alpha_{z_c})}$$

A partir des probabilités d'observer D sous les hypothèses considérées et des distributions a priori, il est possible de calculer les probabilités a posteriori de ces hypothèses sachant D par les formules données précédemment. Une alerte est déclenchée quand les probabilités a posteriori dépassent un seuil pré-déterminé dans un ou plusieurs clusters candidat(s).

Mais l'application de cette méthode nécessite de fixer a priori pour chaque cluster candidat z les paramètres de distribution α_z, β_z, α_{z_c} et β_{z_c} et sa probabilité $Pr(H1(z))$. Il faut aussi fixer les paramètres globaux α_0, β_0 et la probabilité de l'hypothèse nulle $Pr(H0)$.

Si l'on fait l'hypothèse que l'épidémie a autant de chances de se déclarer dans chacune des zones étudiées, on peut fixer le risque global de survenue d'une épidémie dans au moins une de ces zones P_1, et on a alors $P(H0) = 1 - P_1$ et $P(H1(z)) = \frac{P_1}{N_{reg}}$ avec N_{reg} le nombre total de clusters candidats. La valeur de P_1 peut être fixée à dires d'experts, à partir de données historiques ou être utilisé pour contrôler la sensibilité et le risque de faux positifs du système de surveillance.

Les paramètres de la distribution Gamma globale sont eux estimés à partir de l'espérance et la variance de $\frac{x(D,T)}{\mu(D,T)}$, calculées sur données historiques dénuées d'épidémie, de telle

sorte que : $\frac{\alpha_0}{\beta_0} = E\left(\frac{x(D,T)}{\mu(D,T)}\right)$ et que $\frac{\alpha_0}{\beta_0^2} = V\left(\frac{x(D,T)}{\mu(D,T)}\right)$. Ce qui donne :

$$\alpha_0 = \frac{E\left(\frac{x(D,T)}{\mu(D,T)}\right)^2}{V\left(\frac{x(D,T)}{\mu(D,T)}\right)}$$

$$\beta_0 = \frac{E\left(\frac{x(D,T)}{\mu(D,T)}\right)}{V\left(\frac{x(D,T)}{\mu(D,T)}\right)}$$

Les paramètres α_{z_c} et β_{z_c} sont estimés de la même façon à partir de $x(z_c)$ et $\mu(z_c)$, les totaux des comptes et des populations exposées en dehors du cluster candidat z. Pour les paramètres à l'intérieur de chaque cluster candidat sous l'hypothèse alternative, le risque de l'événement n'est plus le même que sous l'hypothèse nulle, il faut donc multiplier l'espérance et la variance de $\frac{x(D,T)}{\mu(D,T)}$ par une estimation du risque relatif p_z, ce qui donne :

$$\alpha_z = p_z \frac{E\left(\frac{x(z)}{\mu(z)}\right)^2}{V\left(\frac{x(z)}{\mu(z)}\right)}$$

$$\beta_z = \frac{E\left(\frac{x(z)}{\mu(z)}\right)}{V\left(\frac{x(z)}{\mu(z)}\right)}$$

En l'absence de connaissance a priori sur le risque relatif attendu, on peut utiliser une distribution uniforme discrétisée de valeurs de p_z, puis faire la moyenne des vraisemblances obtenues. L'auteur recommande de répartir ces valeurs entre 1 et 3 avec un intervalle de 0,2.

Quand aucune donnée historique n'est disponible, l'auteur propose d'utiliser une statistique de scan bayésienne "à l'aveugle". Dans ce contexte on suppose que sous l'hypothèse nulle, pour $\forall(d,t) \in \{D,T\}$, $X(d,t) \sim P(q_0\mu(d,t))$. Ce qui nous donne $E\left(\frac{x}{\mu}\right) = \frac{E(P(q_0\mu))}{\mu} = \frac{q_0\mu}{\mu} = q_0$, et $V\left(\frac{x}{\mu}\right) = \frac{V(P(q_0\mu))}{\mu^2} = \frac{q_0\mu}{\mu^2} = \frac{q_0}{\mu}$. On a donc $\alpha_0 = q_0\mu(D,T)$, $\beta_0 = \mu(D,T)$, $\alpha_{z_c} = q_0\mu(z_c)$, $\beta_{z_c} = \mu(z_c)$, $\alpha_z = p_z q_0 x(z)$ et $\beta_z = \mu(z)$.

3.3 Choix des méthodes de détection

Compte tenu des caractéristiques des méthodes de détection de cluster présentées précédemment, il est possible de sélectionner les méthodes les plus adaptées dans différents contextes. Dans le cadre de cette thèse, l'objectif est de constituer un système de

surveillance épidémiologique passif et orienté vers l'alerte, donc fondé sur la réutilisation de données médicales qui ne seront disponibles que sous forme agrégée. Par ailleurs, la méthode choisie doit permettre de localiser le cluster détecté, et on considère une distribution de Poisson sous l'hypothèse nulle. En complément de ces contraintes, les méthodes envisageables seront comparées sur la base d'une même fenêtre de balayage.

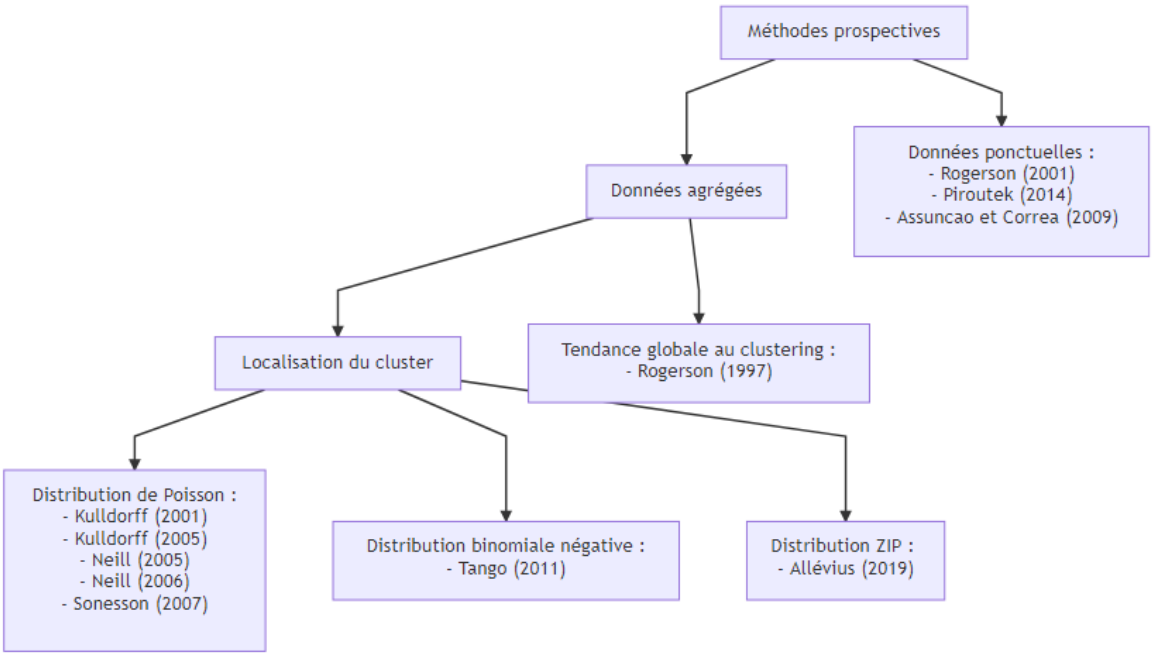

FIG. 5 : Choix des méthodes de détection

On retiendra donc les méthodes suivantes pour les comparer sur simulations :

- méthode de Kulldorff classique (32)
- méthode de Kulldorff par permutations (37)
- méthode non conditionnelle de Poisson (33)
- méthode de CUSUM circulaire (22)
- méthode non conditionnelle binomiale négative (38)

4 Comparaison de méthodes sur simulations

Plusieurs méthodes prospectives de détection de clusters ont été proposées ces 20 dernières années. Certaines sont destinées à des cas d'usages précis, mais dans d'autres cas de figure plusieurs choix sont possibles. C'est par exemple le cas lorsque l'on s'attend à ce que le nombre d'événements observés suivent une distribution de Poisson. La question se pose alors de savoir quelle méthode privilégier dans cette situation. A ce jour, aucune étude n'a évalué les performances de détection de ces méthodes dans des conditions homogènes et sur des critères prenant en compte à la fois la capacité à détecter un cluster, mais aussi la précision spatiale de cette détection.

L'objectif de cette étude de simulation est de comparer les performances des méthodes prospectives de détection de cluster adaptées aux distributions de Poisson sur un même jeu de données simulé. Les indicateurs de performance utilisés devront prendre en compte la capacité à localiser précisément le cluster détecté.

4.1 Données simulées

Pour ce travail on utilisera les données simulées proposées par Kulldorff (2004) pour procéder à des études de comparaison de la puissance de méthodes de détection de cluster (45). Ces données sont présentées de façon complète dans l'article correspondant, et nous ne décrirons ici que les données utilisées dans cette étude.

Les simulations sont produites à l'échelle des 176 codes postaux de la ville de New-York et de la population associée en 2002, soit un total de 8003510 habitants. Les données simulent le nombre de cas d'une pathologie hypothétique observés dans chaque zone associée à un code postal, chaque jour pendant une durée de 31 jours. Elles ont été générées sous l'hypothèse nulle, ainsi que sous plusieurs hypothèses alternatives correspondant à 5 localisations de cluster différentes (nommées A à E), avec à chaque fois une version limitée à une unité spatiale et une autre en recouvrant plusieurs, ainsi que deux niveaux d'augmentation du risque au sein du cluster. Leur localisation peut être visualisée sur la figure 3. Tous les clusters apparaissaient au 31e jour. Sous l'hypothèse nulle, 9999 jeux de données ont été générés, et 1000 pour chaque hypothèse alternative.

Pour chaque jeu de données, le nombre de cas répartis aléatoirement entre les unités était de 3100 (soit 100 multiplié par le nombre de jours). Ce nombre de 100 par jour a été choisi de façon à refléter le nombre de cas observés dans le système de surveillance des urgences de la ville de New-York pour certains syndromes courants. Sous l'hypothèse nulle les cas étaient répartis entre les unités spatiales selon une probabilité dépendant uniquement de la population sous-jacente, c'est-à-dire que le risque était le même pour chaque individu. Sous les hypothèses alternatives, un risque relatif supérieur à un était appliqué aux populations résidant dans le cluster simulé le 31e jour, tandis que la probabilité d'être affecté restait inchangée dans les autres unités et lors des jours précédents, puis les cas étaient répartis aléatoirement selon ces nouvelles probabilités.

Les risques relatifs utilisés sous l'hypothèse alternative diffèrent d'un cluster à l'autre. Ils

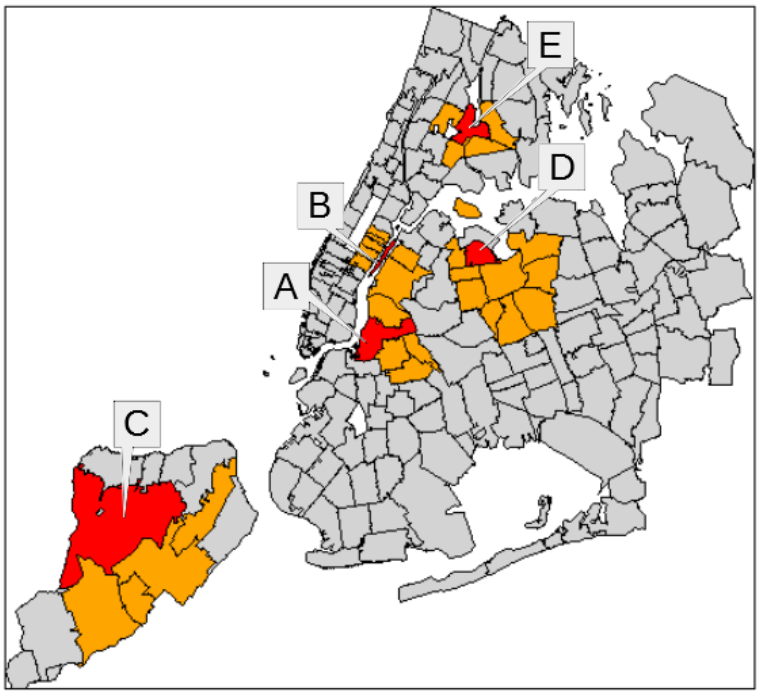

FIG. 6 : Carte de localisation des clusters simulés. Les zones rouges correspondent aux clusters d'une unité spatiale, les clusters étendus couvrent, pour chaque localisation, la zone rouge ainsi que les unités spatiales adjacentes en orange.

ont été attribués de façon à garantir une puissance de 99 % pour un risque de première espèce fixé de 5 %, lorsque l'on cherche à détecter une différence significative entre le risque observé dans les unités spatiales du cluster et le risque sur le reste du territoire étudié, en utilisant une distribution de Poisson, avec des données ne couvrant qu'une seule journée et un total de 100 cas. Il en résulte que le risque relatif associé à chaque épidémie simulée varie entre les différentes localisations adoptées, en fonction de la population sous-jacente. L'objectif de cette méthode est de permettre la comparaison des performances des méthodes en fonction de la taille et de la population du cluster. Ainsi, une différence de puissance observée en fonction des localisations ou tailles de cluster pour une méthode donnée ne résulte pas d'un gain de puissance associé à une population plus importante, mais bien des propriétés de la méthode quand elle est confrontée à différents types de clusters. Pour fournir un risque relatif modéré, la même méthode a été utilisée avec une puissance de 90 %.

Il en résulte les caractéristiques suivantes pour les clusters simulés :

- A (1 unité spatiale (US)) : risque relatif (RR) élevé de 9,91 ; RR modéré de 4,47
- A (5 US) : RR de 4,47
- B (1 US) : RR élevé de 57,08 ; RR modéré non précisé
- B (6 US) : RR de 5,02
- C (1 US) : RR élevé de 9,89 ; RR modéré non précisé
- C (5 US) : RR de 4,93
- D (1 US) : RR élevé de 18,63 ; RR modéré non précisé
- D (10 US) : RR de 3,24
- E (1 US) : RR élevé de 13,76 ; RR modéré non précisé
- E (5 US) : RR de 4,62

4.2 Méthodes et critères de comparaison

Cinq méthodes ont été appliquées sur les jeux de données simulés sous l'hypothèse nulle et sous les différentes hypothèses alternatives, au 31e jour et avec une fenêtre temporelle maximale de 3 jours. Il s'agit des méthodes de Kulldorff classique et par permutations, de la méthode non conditionnelle pour distribution de Poisson de Neill, de la méthode non conditionnelle pour distribution binomiale négative de Tango, et de la CUSUM circulaire de Sonesson. Pour chaque méthode, seul le cluster le plus probable a été étudié. La phase d'inférence était réalisée avec 999 simulations de Monte Carlo. Les analyses étant réalisées à un seul temps t, il n'y a pas eu de prise en compte de tests multiples.

Pour la méthode non conditionnelle de Neill et la CUSUM circulaire de Sonesson, le nombre de cas attendus a été calculé en appliquant le taux d'événements observé dans l'ensemble de la population sur les 28 premiers jours à la population associée à chaque code postal pour les jours 29 à 31, dans chaque jeu de données simulé. Pour la méthode non conditionnelle de Tango, une régression binomiale négative était réalisée avec les nombres de cas observés au cours des 28 premiers jours par code postal en variable à expliquer, et les populations associées à ces codes postaux en variable explicative, dans chaque jeu de données simulé. Puis, les nombres de cas attendus étaient prédits à partir

des résultats du modèle et des populations de chaque unité spatiale.

Sous l'hypothèse nulle, l'objectif était d'évaluer le contrôle du risque de première espèce de conclure à tort à l'existence d'un cluster alors qu'il n'y en avait pas. Les analyses ont été réalisées au risque α nominal de 5 %, et un risque de première espèce empirique a été calculé à partir des résultats obtenus sur les 9999 simulations pour évaluer le degré de contrôle réel de ce risque.

Sous les hypothèses alternatives, l'objectif était de comparer les méthodes sur leur capacité à détecter le véritable cluster. Les analyses ont été réalisées au risque α nominal de 5 %, et différents indicateurs de comparaison se basant sur un calcul de puissance ont été considérés (35,46) :

- La puissance standard
- Les indicateurs étendus : prise en compte de la précision spatiale de la détection
  - puissance étendue (spatiale ou spatio-temporelle)
  - sensibilité
  - valeur prédictive positive

La puissance standard est la probabilité de rejet de l'hypothèse nulle avec un risque de première espèce de 5 %, indépendamment de la localisation du cluster ayant conduit au rejet. C'est-à-dire qu'on ne vérifie pas la concordance entre le cluster détecté et le cluster réel. Si on l'évalue sur un nombre n de jeux de données simulés sous l'hypothèse alternative, la puissance est alors la proportion de ces n jeux de données pour lesquels il y a rejet de l'hypothèse nulle.

Un défaut inhérent à la puissance standard est l'absence de prise en compte de la dimension spatiale du processus de détection, c'est pourquoi elle n'a pas été retenue dans le cadre de ce travail. D'autres indicateurs ont été proposés afin de corriger ce défaut. Pour cela ils s'appuient sur la distribution de puissance spatiale proposée par Takahashi et al. (2006) (46) et son extension spatio-temporelle proposée par Takahashi et al. (2008) (35).

La distribution de puissance spatiale $P_0(l, s|s^*)$ est la distribution bivariée correspondant à la probabilité d'obtenir un cluster significatif de taille l et recoupant le vrai cluster de s unités spatiales, sachant que ce dernier est de taille s^* (46). On a donc $l \geq 1, 0 \leq s \leq s^*$ et $P_0(l, s|s^*) = Pr\{L = l, S = s|s^*\}$ avec L et S les variables aléatoires associées à l, et s.

De la même façon il est possible de définir une version spatio-temporelle trivariée de la puissance telle que $P_0(l, s, t|s^*, t^*) = Pr\{L = l, S = s, T = t|s^*, t^*\}$, avec t la durée du cluster significatif le plus probable ($1 \leq t \leq t_{max}$), T la variable aléatoire associée et t^* la durée du véritable cluster (35).

Les faux négatifs spatiaux (FN_s) correspondent aux unités spatiales faisant partie du cluster réel, mais pas du cluster significatif le plus probable, soit $FN_s = s^* - s$. Les faux positifs spatiaux (FP_s) correspondent aux unités spatiales faisant partie du cluster significatif le plus probable, mais pas du cluster réel, soit $FP_s = l - s$. De la même façon on peut définir des faux positifs (FP_t) et des faux négatifs temporels (FN_t).

A partir de la distribution de puissance spatio-temporelle on peut calculer une puissance spatio-temporelle comme le nombre de simulations dans lesquelles la proportion du cluster réel qui a été détectée par la méthode dépasse un seuil donné. On peut fixer un seuil différent dans le domaine temporel et dans le domaine spatial, et le faire varier afin de comparer les performances des méthodes pour différents seuils. Dans cette étude, le cluster venant d'apparaitre n'a qu'une durée de 1 jour, donc toute détection de cluster entraine nécessairement un recoupement de 100 % du cluster réel dans le domaine purement temporel. On peut alors simplifier l'indicateur en ne considérant qu'un seul seuil, pour la dimension spatiale, et que l'on fera varier entre 0 et 100 % de recoupement lorsque le cluster réel est constitué de plusieurs unités spatiales.

Takahashi et al. (2008) (35) proposent aussi d'autres indicateurs se basant sur la distribution spatiale de puissance pour la comparaison de méthodes de détection de cluster : la sensibilité et la valeur prédictive positive. Ces indicateurs peuvent être définis en nombre d'unités spatiales ou en population.

La sensibilité (Se_1) est la probabilité d'identifier les unités spatiales faisant réellement partie du cluster. Cette valeur est donnée par l'espérance de $\frac{s}{s^*}$, soit :

$$Se_1 = E\left(\frac{S}{s^*}\right) = \sum_{s=0}^{s^*} \frac{s}{s^*} P_0(+, s|s^*)$$

où le signe + désigne l'ensemble des situations où s ou l sont positifs et $P_0(+, s|s^*)$ la probabilité d'obtenir un cluster significatif partageant s unités spatiales avec le cluster réel de taille s^*. Son équivalent en contexte populationnel Se_2 est donné par le rapport entre l'espérance de la population totale des S unités correctement identifiées, et la population totale des s^* unités du cluster réel. On peut aussi calculer son équivalent spatio-temporel par l'espérance de $\frac{s}{s^*}$ pondérée par $\frac{t}{t^*}$, soit :

$$Se_1 = E\left(\frac{S}{s^*}\frac{T}{t^*}\right) = \sum_{s=0}^{s^*}\sum_{t=0}^{t^*} \frac{s}{s^*}\frac{t}{t^*} P_0(+, s, t|s^*, t^*)$$

La valeur prédictive positive (VPP_1) se définit comme la probabilité qu'une unité faisant partie du cluster significatif le plus probable fasse bien partie du cluster réel. Elle est donnée par l'espérance de $\frac{s}{l}$ ($l > 0$), soit :

$$VPP_1 = E\left(\frac{S}{L}\middle| L > 0\right) = \sum_{l\geq 1}\sum_{s\geq 0} \frac{s}{l}\frac{P_0(l, s|s^*)}{P_0(+, +|s^*)}$$

où le signe + désigne l'ensemble des situations où s ou l sont positifs et $\frac{P_0(l,s|s^*)}{P_0(+,+|s^*)}$ la probabilité de détecter un cluster de taille l partageant s unités spatiales avec le cluster

réel de taille s^*, sachant que l'on a détecté un cluster. Son équivalent en contexte populationnel VPP_2 est donné par l'espérance du rapport entre la population totale des S unités spatiales qui font vraiment partie du cluster, et celle de l'ensemble des L unités du cluster significatif le plus probable. Soit $s_t = \min(t, t^*)$ la durée du cluster détecté qui est partagée avec le cluster réel ; on peut aussi calculer l'équivalent spatio-temporel de VPP_1 par l'espérance de $\frac{s}{l}$ ($l > 0$), pondérée par $\frac{s_t}{t}$, soit :

$$VPP_1 = E\left(\frac{S}{L}\frac{S_t}{T}\middle| L > 0\right) = \sum_{l\geq 1}\sum_{s\geq 0}\sum_{s_t=0}^{t^*}\frac{s}{l}\frac{s_t}{t}\frac{P_0(l,s,t|s^*,t^*)}{P_0(+,+,+|s^*,t^*)}$$

Une autre façon de comparer les méthodes de détection de cluster dans le cadre prospectif est l'étude du délai jusqu'à détection du cluster lors d'analyses itératives au cours du temps, l'objectif étant bien entendu de minimiser ce délai. Plusieurs indicateurs ont été proposés dans ce cadre (47,48) toutefois, comme la puissance standard, ils ont le défaut de ne pas prendre en compte la dimension spatiale de la détection (40). Ils ne sont donc pas détaillés ici.

Dans le cadre de ce travail, l'objectif était de comparer la capacité des différentes méthodes évaluées à détecter un cluster, en prenant en compte la précision spatiale de cette détection. Les indicateurs retenus pour la comparaison des méthodes étaient donc :

- La puissance spatio-temporelle avec un seuil de recoupement spatial variable de 0 à 100% lorsque le cluster comportait plusieurs unités, et un seuil fixe de 50 % quand on souhaitait considérer une seule valeur de puissance ou quand le cluster ne était une unique unité spatiale
- La sensibilité spatio-temporelle
- La valeur prédictive positive spatio-temporelle

Les analyses ont été réalisées avec le logiciel R version 3.6.1 (2019-07-05) et le package scanstatistics version 1.0.1 (49).

4.3 Résultats

Lors des simulations sous l'hypothèse nulle, la méthode de Kulldorff obtient un risque α empirique de 5.17 % [4.74 ; 5.62], la méthode par permutations de 7.12 % [6.62 ; 7.64], la méthode de Neill de 3.76 % [3.4 ; 4.15], la méthode de Tango de 2.21 % [1.93 ; 2.52] et la C-CUSUM de Sonesson de 5.61 % [5.17 ; 6.08].

Concernant la puissance, les méthodes de Kulldorff classique et de Neill présentent des résultats très similaires quelle que soit la taille du cluster ou l'importance du risque relatif (RR) associé. La méthode statistique de scan par permutations obtient des résultats comparables aux deux précédentes quand le cluster consiste en une seule unité spatiale avec un RR élevé, mais elle est un peu meilleure lorsque le risque est modéré, et un peu plus faible quand le cluster regroupe plusieurs unités spatiales. La méthode de Tango obtient des résultats très variables en fonction de l'importance du RR associé au cluster : sa

Tab. 1 : Puissance spatio-temporelle obtenue par simulation

Cluster	Population*	RR	Kulldorff	Permutations	Neill	Tango	C-CUSUM
A	1.06	élevé	0.852	0.844	0.852	0.217	0.644
B	0.12	élevé	0.910	0.905	0.909	0.960	0.397
C	1.06	élevé	0.848	0.839	0.852	0.214	0.595
D	0.45	élevé	0.848	0.857	0.851	0.921	0.532
E	0.67	élevé	0.829	0.838	0.828	0.799	0.542
A	1.06	modéré	0.310	0.322	0.305	0.008	0.179
B	0.12	modéré	0.346	0.362	0.340	0.532	0.067
C	1.06	modéré	0.303	0.305	0.297	0.007	0.111
D	0.45	modéré	0.300	0.336	0.286	0.424	0.112
E	0.67	modéré	0.262	0.334	0.266	0.220	0.134
A (5 US)	3.98		0.774	0.728	0.761	0.000	0.788
B (6 US)	3.20		0.755	0.703	0.746	0.000	0.782
C (5 US)	3.31		0.774	0.746	0.775	0.000	0.710
D (10 US)	8.06		0.782	0.715	0.761	0.000	0.876
E (5 US)	3.73		0.799	0.756	0.790	0.000	0.771

RR = Risque Relatif

* En pourcentage de la population totale

puissance est très supérieure aux autres méthodes quand il est très élevé (clusters B et D), mais elle diminue très rapidement voire s'annule lorsqu'il diminue. Enfin, la C-CUSUM obtient également des résultats très variables en fonction du cluster étudié, les meilleures performances étant obtenues avec le cluster A et les plus mauvaises avec le B. A cluster identique (même unité spatiale), les performances augmentent avec le risque relatif au sein du cluster, et la méthode se révèle plus puissante que les autres lorsque les clusters sont de grande taille.

Le critère de la sensibilité présente des résultats très comparables à ceux de la puissance, avec des méthodes de Kulldorff classique et de Neill aux performances similaires, et une méthode par permutations qui obtient des résultats approximativement équivalents aux deux précédentes, mais un peu meilleurs pour les RR modérés et un peu plus faibles pour les clusters de plusieurs unités spatiales. Là encore, la méthode de Tango est très dépendante du RR, avec une sensibilité supérieure aux autres méthodes lorsqu'il est très élevé, mais qui diminue très rapidement avec lui. La C-CUSUM obtient avec ce critère des résultats très comparables à ceux obtenus avec la puissance : une variabilité importante d'un cluster à l'autre, mais une sensibilité qui augmente avec le risque relatif à cluster identique. Cette méthode est également la plus sensible sur les clusters de grande taille.

Concernant la valeur prédictive positive, les méthodes de Neill et par permutations obtiennent des résultats globalement équivalents et systématiquement un peu supérieurs à ceux de la méthode de Kulldorff classique. La méthode de Tango obtient les meilleures

Tab. 2 : Sensibilité spatio-temporelle obtenue par simulation

Cluster	Population*	RR	Kulldorff	Permutations	Neill	Tango	C-CUSUM
A	1.06	élevé	0.852	0.844	0.852	0.217	0.644
B	0.12	élevé	0.910	0.905	0.909	0.960	0.397
C	1.06	élevé	0.848	0.839	0.852	0.214	0.595
D	0.45	élevé	0.848	0.857	0.851	0.921	0.532
E	0.67	élevé	0.829	0.838	0.828	0.799	0.542
A	1.06	modéré	0.310	0.322	0.305	0.008	0.179
B	0.12	modéré	0.346	0.362	0.340	0.532	0.067
C	1.06	modéré	0.303	0.305	0.297	0.007	0.111
D	0.45	modéré	0.300	0.336	0.286	0.424	0.112
E	0.67	modéré	0.262	0.334	0.266	0.220	0.134
A (5 US)	3.98		0.698	0.658	0.680	0.019	0.756
B (6 US)	3.20		0.687	0.642	0.675	0.036	0.734
C (5 US)	3.31		0.720	0.689	0.718	0.045	0.693
D (10 US)	8.06		0.717	0.655	0.695	0.006	0.816
E (5 US)	3.73		0.748	0.708	0.738	0.025	0.744

RR = Risque Relatif

* En pourcentage de la population totale

performances dans la plupart des cas (risques relatifs élevés ou clusters de plusieurs unités spatiales), mais elle est faible pour les clusters localisés ayant le plus faible RR (clusters A et C avec risque modéré). La méthode C-CUSUM est celle qui obtient les plus faibles valeurs prédictives positives, intégrant dans tous les cas dans son cluster le plus probable davantage d'unités spatiales qui ne font pas partie du cluster réel que d'unités qui en font effectivement partie. Toutefois on peut observer que, comme avec les indicateurs précédents, la VPP est meilleure avec les clusters A et C, et moins bonne avec le B, qu'elle s'améliore avec l'augmentation du RR, et qu'elle est nettement plus élevée pour les clusters de grande taille.

La figure 7 présente l'évolution de la puissance spatiale en fonction du seuil de recouvrement du cluster réel adopté pour chaque cluster multi-unités et chaque méthode utilisée. Un seuil de 0,5 indique qu'au moins 50 % de ses unités spatiales doivent faire partie du cluster le plus probable, tandis qu'un seuil de 1 indique que l'intégralité de ses unités spatiales doivent avoir été détectées. Les méthodes de Kulldorff classique, de Neill et par permutations présentent des profils très similaires et relativement stables face à l'évolution du seuil, mais des baisses de puissance plus importantes peuvent être observées quand le seuil dépasse les 70 % de recouvrement. La méthode de Kulldorff parait toutefois systématiquement un peu supérieure et la méthode par permutations un peu inférieure. La méthode de Tango, conformément aux observations précédentes, obtient des résultats nuls pour la plupart des seuils testés, mais quand elle détecte un cluster, celui-ci

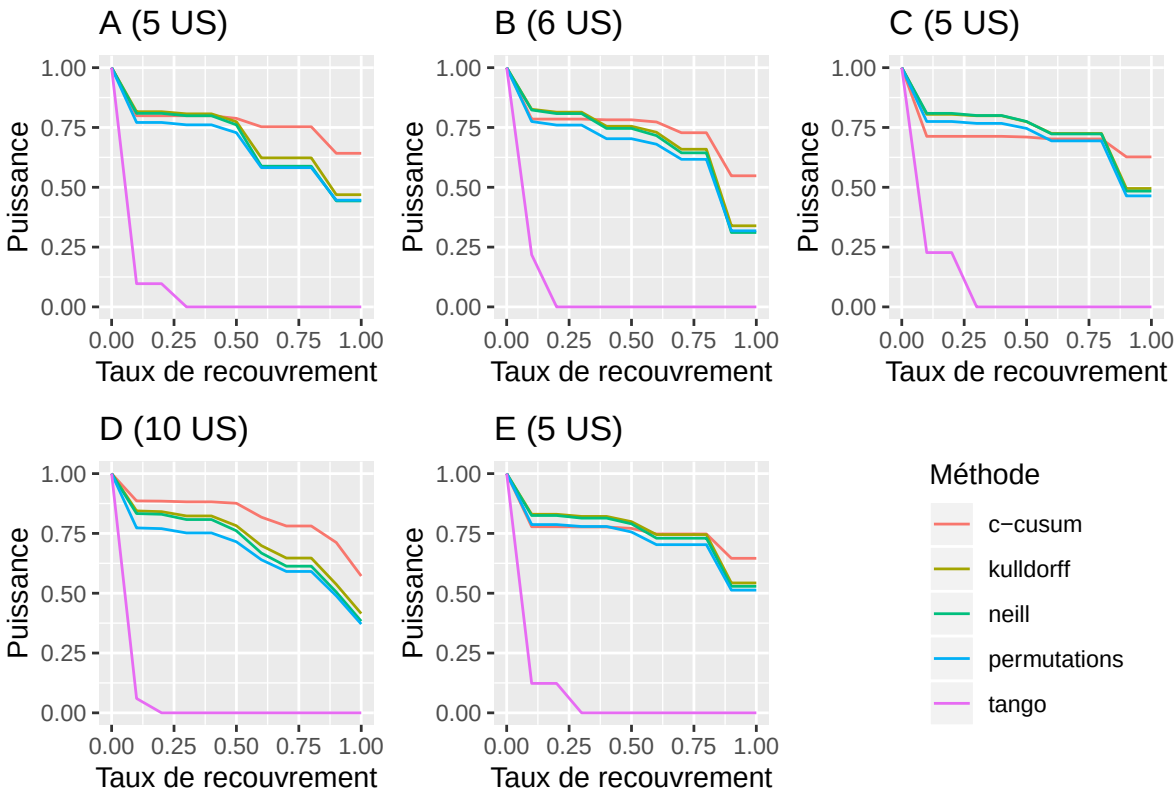

FIG. 7 : Evolution de la puissance spatiale en fonction du seuil de taux de recouvrement du cluster réel par le cluster détecté. Ne sont représentés que les cluster consitués de plusieurs unités spatiales (US).

TAB. 3 : Valeur prédictive positive spatio-temporelle obtenue par simulation

Cluster	Population*	RR	Kulldorff	Permutations	Neill	Tango	C-CUSUM
A	1.06	élevé	0.879	0.900	0.893	0.956	0.137
B	0.12	élevé	0.975	0.979	0.979	1.000	0.083
C	1.06	élevé	0.846	0.842	0.853	0.926	0.193
D	0.45	élevé	0.954	0.958	0.961	0.999	0.117
E	0.67	élevé	0.928	0.943	0.940	0.991	0.133
A	1.06	modéré	0.686	0.727	0.731	0.348	0.074
B	0.12	modéré	0.830	0.832	0.862	0.973	0.026
C	1.06	modéré	0.644	0.672	0.670	0.219	0.084
D	0.45	modéré	0.800	0.817	0.833	0.955	0.050
E	0.67	modéré	0.755	0.771	0.787	0.913	0.064
A (5 US)	3.98		0.697	0.715	0.737	0.836	0.342
B (6 US)	3.20		0.792	0.811	0.822	0.924	0.359
C (5 US)	3.31		0.722	0.737	0.732	0.938	0.428
D (10 US)	8.06		0.707	0.726	0.751	0.769	0.461
E (5 US)	3.73		0.828	0.856	0.854	0.885	0.431

RR = Risque Relatif

* En pourcentage de la population totale

recouvre presque toujours une part minoritaire des unités spatiales du cluster réel. Enfin, la C-CUSUM présente le profil le plus régulier, c'est-à-dire que sa puissance est moins susceptible de diminuer quand le seuil de recouvrement adopté augmente. Par ailleurs, elle est la méthode la plus puissante dans la plupart des cas et elle l'est systématiquement lorsque l'on souhaite que l'intégralité du cluster réel soit détecté.

4.4 Discussion

Dans le cadre de ce travail de simulations, cinq méthodes de détection de clusters ont été comparées. Celles qui apparaissent les plus stables en termes de puissance et de sensibilité sont les méthodes conditionnelle et non conditionnelle pour distribution de Poisson, ainsi que la méthode de scan par permutations, avec un léger avantage aux deux premières. Cette stabilité témoigne d'une relativement faible dépendance à la localisation et la taille du cluster, ainsi qu'à sa population. On peut aussi noter que la méthode non conditionnelle de Neill, parait un peu moins puissante mais plus spécifique, avec une meilleure VPP que les deux autres. Enfin, la méthode par permutations semble moins bien contrôler le risque de première espèce, avec un α empirique de 7,12 % là où la méthode conditionnelle de Kulldorff est très proche des 5 % attendus, et la méthode de Neill un peu plus conservatrice à 3,76 %.

Lors de la proposition de la méthode de statistique de scan par permutations, Kulldorff

n'avait pas comparé ses performances à celles de la statistique de scan conditionnelle de Poisson. En 2009, Neill avait comparé sa méthode non conditionnelle à celle de Kulldorff sur un critère de délai de détection (16). Il avait alors conclu qu'elle identifiait les clusters légèrement plus rapidement lorsqu'ils étaient de taille petite à moyenne (< 10 % et 10-25 % du territoire étudié respectivement), mais était jusqu'à 5 fois plus rapide lorsqu'ils couvraient l'intégralité du territoire. Ce résultat était attendu dans la mesure où la méthode de Kulldorff compare le risque à l'intérieur du cluster au risque à l'extérieur, alors que celle de Neill ne fait qu'évaluer si le risque est plus important qu'attendu au sein du cluster candidat. Les résultats de ce travail de thèse ne permettent pas de mettre en évidence un écart de puissance significatif entre les deux méthodes, mais il faut noter que les clusters utilisés sont de taille plus modeste que dans l'étude de Neill, représentant ici moins de 10 voire moins de 5 % du territoire étudié et de sa population. Les résultats sont donc compatibles avec ceux obtenus par Neill sur les clusters de petite taille. Dans la même étude, Neill obtenait également une précision spatiale un peu meilleure avec sa méthode, ce qui est cohérent avec les résultats présentés ici, où la sensibilité des deux méthodes est quasiment équivalente mais la VPP sensiblement plus élevée avec la méthode non conditionnelle.

La méthode de Tango semble la plus conservatrice de toutes les méthodes évaluées, avec un risque α empirique plus de deux fois inférieur aux 5 % attendus. Elle présente aussi des performances en termes de sensibilité et de puissance extrêmement variables en fonction du risque relatif et de la taille du cluster. En revanche sa valeur prédictive positive est presque toujours la plus élevée avec des valeurs souvent comprises entre 90 et 100 %, ce qui indique que la faible sensibilité est contrebalancée par une très forte spécificité.

Dans l'article présentant sa méthode, Tango l'a comparée à la méthode conditionnelle de Kulldorff sur la base de critères construits à partir de la distribution spatio-temporelle de puissance (38). Ces critères étaient la puissance standard, la probabilité de détection exacte, la probabilité que les régions détectées fassent toutes partie du cluster réel, et la probabilité qu'il y ait au plus deux unités spatiales erronées dans le cluster détecté, ils sont donc principalement centrés sur la spécificité. La puissance standard était presque systématiquement plus faible (souvent de façon très importante) avec la méthode de Tango, sauf dans quelques situations où le risque relatif était un peu plus élevé, mais plus faible que dans les simulations utilisées pour cette thèse. A contrario, sur les trois autres critères la méthode non conditionnelle binomiale négative est presque toujours supérieure à la méthode conditionnelle, elle est donc plus spécifique. Les résultats obtenus dans ce travail de thèse sont donc compatibles avec ceux de Tango. Sa méthode vise à ne détecter que des unités spatiales faisant partie du cluster réel, quitte à n'en détecter qu'une partie ou à ne pas détecter de cluster du tout lorsque ce n'est pas possible.

D'autres facteurs pourraient aussi contribuer à expliquer les performances de la méthode de Tango. En effet, cette méthode est conçue pour des distributions binomiales négatives, et les simulations de Tango avaient été réalisées sur ce type de données, alors que les données simulées utilisées dans ce travail suivaient une distribution de Poisson. Toutefois, cette méthode est censée pouvoir s'adapter à une distribution de Poisson, la distribution binomiale négative permettant simplement de prendre en compte les situations de surdis-

persion lorsqu'elles sont présentes. Par ailleurs, l'estimation des nombres de cas attendus n'a pu être réalisée indépendamment dans chaque unité spatiale comme le propose Tango dans son article, car le faible taux d'événements simulé par Kulldorff impliquait l'existence de nombreuses unités n'ayant connu aucun cas sur la période d'apprentissage, ce qui aurait conduit à un effectif attendu nul, valeur qui n'est pas autorisée par la méthode de calcul de la statistique. On ne peut donc exclure une perte de précision dans l'estimation des attendus ayant pu impacter les performances de la méthode.

Les performances de la C-CUSUM peuvent paraître paradoxales, dans la mesure où elles augmentent avec le risque relatif à cluster identique, alors qu'elles semblent évoluer en sens inverse lorsque l'on compare les clusters A à E entre eux (le B ayant à chaque fois le RR le plus élevé, et le A le plus faible). Toutefois, la méthode de simulation adoptée par Kulldorff fait qu'à catégorie de risque identique (modéré ou élevée), le risque relatif associé à chaque cluster est dépendant de sa population : plus la population est faible, plus le risque est élevé. Cette méthode compare les nombres de cas observés aux moyennes attendues sous les hypothèses nulle (λ_0) et alternative (λ_1) pour la loi de Poisson, or ces moyennes dépendent à la fois du risque de survenue de l'événement et de la population sous-jacente. Les performances obtenues par cette méthode augmentent donc avec le RR (*cf* résultats à cluster identique), mais elles augmentent aussi fortement avec le nombre de cas attendus, et donc la population des unités spatiales. Cette dépendance expliquerait aussi pourquoi elle obtient de meilleurs résultats lorsque les clusters sont de grande taille. Par ailleurs, on peut remarquer que la faible VPP obtenue en conjonction avec la très bonne couverture spatiale des clusters multi-unités semble indiquer que cette méthode a tendance à détecter des clusters de grande taille, plus grands qu'ils ne le sont en réalité.

Cette méthode avait été comparée à celle de Kulldorff par Sonesson (22), qui concluait à la supériorité de la C-CUSUM sur un critère de délai de détection. Toutefois, les simulations de cette étude portaient uniquement sur des populations réparties de façon homogène entre les unités spatiales. Des travaux supplémentaires seront nécessaires pour étudier plus précisément l'impact de la population des unités spatiales sur les performances de la C-CUSUM. Mais les résultats des simulations menées dans le cadre de ce travail de thèse suggèrent que cette méthode puisse être plus puissante que les statistiques de scan dans certaines configurations (clusters de grande taille, nombre de cas attendus important), d'autres études devront permettre de mieux cerner ces limites. En l'état, il n'est pas encore possible de fixer précisément les conditions dans lesquelles cette méthode pourrait être préférable à une autre.

En conclusion, compte tenu de sa nature non conditionnelle et des résultats de cette étude, on peut recommender en première intention dans le cas prospectif l'utilisation de la méthode non conditionnelle pour distribution de Poisson proposée par Neill. Toutefois, lorsqu'il n'est pas possible d'estimer de façon fiable le nombre de cas attendus, par exemple dans le cas d'une pathologie émergente ou en l'absence de données historiques concernant son incidence, la méthode conditionnelle de Kulldorff peut s'avérer une bonne alternative car se basant plutôt sur une proportionnalité du risque par rapport à la population exposée. Quand il est impossible d'estimer la population à risque, la méthode de statistique de scan par permutations permet d'obtenir une puissance et une valeur prédictive

positive presque équivalentes aux deux précédentes. La méthode non conditionnelle pour distribution binomiale négative de Tango pourrait être intéressante dans les cas où on privilégie avant tout la spécificité. Toutefois une très faible puissance est obtenue avec les plus faibles risques relatifs et avec les clusters de grande taille, sans qu'il soit possible de déterminer si seul le premier facteur ou les deux ont une influence sur ses performances. Par ailleurs, dans le cadre de la surveillance épidémiologique orientée vers l'alerte, la sensibilité est généralement privilégiée sur la spécificité, afin de ne pas manquer un événement sanitaire important, qui plus est des étapes de validation du signal permettent d'éliminer *a posteriori* les faux positifs. Des études supplémentaires seraient utiles pour mieux cerner les limites de cette méthode. Enfin, la C-CUSUM de Sonesson parait intéressante dans les situations où le nombre de cas attendus est important où si l'on s'attend à des clusters de grande taille, mais là encore des études supplémentaires devront permettre de mieux identifier ses cas d'usage.

5 Applications en épidémiologie

Dans la première partie de cette thèse, le concept de surveillance épidémiologique a été présenté pour aboutir au constat de l'existence d'un besoin en matière de détection de territoires nécessitant une intervention de santé publique. Dans la seconde partie, les différentes méthodes de détection de cluster d'événements de santé pouvant être utilisés dans le cadre de la surveillance épidémiologique, donc de nature prospective, ont été présentées. Après prise en compte des contraintes associées à la réutilisation de données médicales, plusieurs méthodes de détection étaient envisageables. La troisième partie a alors permis de comparer leurs performances sur des données simulées pour aboutir à une priorisation des méthodes à utiliser selon le contexte.

Il s'agit désormais d'évaluer l'intérêt des résultats obtenus par une telle méthode lorsqu'elle est appliquée dans des contextes épidémiologiques réels. Pour cela, quatre situations seront testées :

- la détection d'une épidémie de maladie connue, en l'ocurrence dans le cadre de la surveillance épidémiologique de la grippe saisonnière à partir de données de consultations aux urgences ;
- la détection de clusters de diffusion active d'une pathologie infectieuse émergente : la maladie à coronavirus de 2019 ;
- l'évaluation de la gravité de situations sanitaires exceptionnelles par l'identification d'épisodes de surmortalité ;
- le ciblage de territoires pour le déploiement d'actions de prévention / promotion de la santé, ici pour les besoins en matière de contraception et prévention des grossesses non désirées.

5.1 Consultations aux urgences pour grippe

La grippe est une pathologie virale respiratoire provoquant des épidémies saisonnières. En France, ces épidémies durent de 5 à 16 semaines, avec une moyenne de 9,3 et ont habituellement lieu entre les mois de décembre et avril (50,51). Elles touchent en moyenne 2,5 millions de personnes et provoquent de 8000 à 15000 morts (50). C'est cet impact sanitaire, mais aussi ses conséquences socio-économiques qui justifient la surveillance régulière et en temps réel de la grippe. Les objectifs de cette surveillance couvrent entre autres la détection précoce de la survenue des épidémies, le suivi de leur évolution et de leurs conséquences sanitaires notamment en termes de formes graves et de décès (50). La détection précoce permet en particulier la mise en oeuvre de campagnes de prévention passant par la promotion des gestes barrières et de la vaccination.

En France, la surveillance épidémiologique de la grippe est coordonnée par l'agence nationale Santé Publique France, et se base sur différents types de données (50) :

- en médecine ambulatoire : réseau Sentinelles pour la médecine générale, associations SOS médecins

- à l'hôpital : réseau OSCOUR de surveillance des services d'urgences, hospitalisations en réanimation
- en établissement d'hébergement pour personnées âgées : cas groupés d'infections respiratoires aiguës
- surveillance virologique du Centre National de Référence des virus causant des infections respiratoires
- données de mortalité

Toutefois ce système repose, pour l'identification des épidémies, sur des méthodes de suivi temporel par méthode de Shewhart à l'échelle régionale (51), or Neill a montré que le suivi temporel de plusieurs unités spatiales en parallèle est moins performant que les méthodes de détection de cluster pour l'identification d'épidémies, surtout lorsqu'elles concernent plusieurs territoires simultanément (16). Ces méthodes pourraient donc permettre d'identifier plus rapidement leur survenue.

L'objectif de ce travail était de pouvoir détecter l'apparition de l'épidémie de grippe saisonnière à partir des données de consultation pour grippe aux urgences.

5.1.1 Matériel et méthodes

Les données provenaient de l'observatoire national des urgences français OSCOUR et mises à disposition via la plate-forme GEODES (Géo données en santé publique) de Santé Publique France. Elles consistaient en des taux de passage aux urgences pour grippe par département et par semaine de 2010 à 2019. Le nombre total de passage aux urgences hebdomadaire et par département sur la même période étant indisponible, il a été estimé à partir de la Statistique Annuelle des Etablissements (SAE). La SAE fourni le nombre total de consultations aux urgences par établissement sur l'année étudiée de 2013 à 2018. Ces effectifs ont été agrégés par département afin d'obtenir le nombre annuel de consultations aux urgences par département et sa moyenne hebdomadaire. En l'absence de données de la SAE pour les années 2010-2012 et 2019, elles ont été retirées de l'analyse.

Une période d'apprentissage de 2013 à 2016 était utilisée pour le calcul du nombre de cas attendus au cours de la période d'étude, soit 2017-2018. Lorsque le taux observé dans un département était manquant pendant la période d'apprentissage, il était imputé en réalisant la moyenne des taux des départements limitrophes au cours de la même semaine. Lorsque le taux observé était manquant pendant la période d'étude, il était remplacé par le taux attendu.

Pour chaque département et chaque semaine, le nombre de cas observés était estimé par l'application du taux de consultation pour grippe (pour 10000 passages aux urgences) au nombre hebdomadaire moyen de consultations aux urgences réalisées la même année et était arrondi à l'unité. Le taux de consultations attendu était obtenu dans chaque département par une régression périodique de Poisson incluant une tendance temporelle linéaire, une périodicité annuelle et une périodicité semestrielle, paramétrage adopté dans le suivi de l'incidence de la grippe en France depuis plusieurs décennies (52). Le modèle obtenu était alors :

$$\ln(Y_t) = a_0 + a_1 t + \gamma_1 \cos(2\pi t/n) + \delta_1 \sin(2\pi t/n) + \\ \gamma_2 \cos(4\pi t/n) + \delta_2 \sin(4\pi t/n) + \epsilon_t$$

avec Y_t la variable à expliquer (le taux de consultation pour grippe pour 10000 passages aux urgences), a_0 l'intercept, a_1, γ et δ les coefficients associés aux variables explicatives, t le temps en semaines, n la durée d'une période (soit 52,179 pour une année) et ϵ_t l'erreur résiduelle. Aucun offset n'a été utilisé car le taux se rapporte toujours à 10000 consultations à chaque temps t, l'offset n'aurait donc consisté qu'en l'ajout d'une constante au modèle. Le nombre de cas attendu était ensuite obtenu en appliquant le taux attendu au nombre hebdomadaire moyen de consultations aux urgences pour le département et l'année considérés.

Des cartes de ratio d'incidence ont été produites pour chaque date étudiée (en 2017-2018) à partir des effectifs observés et attendus calculés précédemment.

Une analyse de détection de cluster prospective a été réalisée avec la méthode non conditionnelle pour distribution de Poisson (33). La fenêtre de balayage utilisée était circulaire selon la méthode de Kulldorff (32). Aucun seuil de population ou de taille maximale n'a été imposé. En effet, étant donné que la méthode de Neill ne compare pas l'intérieur et l'extérieur du cluster, détecter un cluster à un endroit n'équivaut pas à détecter un cluster de risque diminué ailleurs. Par ailleurs, l'épidémie de grippe saisonnière finit généralement par couvrir l'ensemble du territoire national pendant une partie de sa durée, il était donc important de pouvoir le prendre en compte. Les analyses débutaient en janvier 2017 et se terminaient en décembre 2018. Chaque analyse portait sur le mois en cours ainsi que tous les mois précédents depuis le début du processus (janvier 2017).

Le risque de faux positif associé aux situations de tests multiples a été contrôlé par la méthode de Kulldorff (32,42) de façon à garantir un risque de première espèce de 5 % sur 52 analyses, soit une année de suivi. Elle consistait à prendre en compte les simulations de Monte Carlo réalisées pour les analyses précédentes lors de la phase d'inférence.

Enfin, des cartes étaient produites pour visualiser l'ensemble des clusters détectés par la méthode précédente.

Les analyses statistiques ont été réalisées avec le logiciel R version 3.6.1 (2019-07-05) et le package scanstatistics version 1.0.1 (49).

5.1.2 Résultats

Les cartes de ratio d'incidence ne sont présentées que pour quelques semaines réparties sur l'ensemble de la période étudiée. Elles permettent très clairement d'identifier les épidémies de grippe des hivers 2016-17 et 2017-18, qui à leur apogée couvrent l'ensemble du pays. Au printemps et pendant l'été les cartes témoignent d'une très faible incidence, puis à partir de l'automne divers foyers localisés de grippe apparaissent de façon ponctuelle et plus ou moins temporaire avant une diffusion plus large et le début de l'épidémie de grippe suivante.

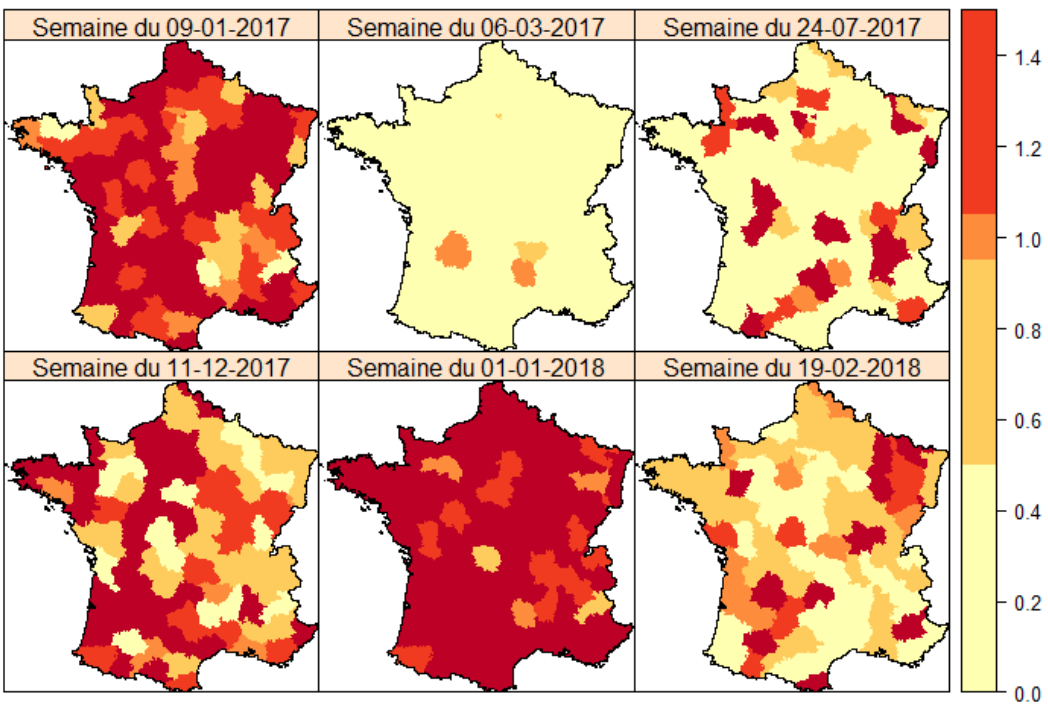

FIG. 8 : Cartes de ratio d'incidence de consultation pour grippe aux urgences à différents temps de l'analyse. Les dates indiquées correspondent au premier jour de la semaine.

Dès la première semaine d'analyse, un cluster est détecté couvrant la plupart du pays à l'exception de l'ensemble de sa côte Ouest. En semaine deux, il se généralise à la quasi-totalité de la métropole avant de régresser pour ne plus couvrir que le Sud-Ouest de la semaine 3 à la semaine 5. De la semaine 6 à la semaine 49 on observe un cluster significatif correspondant au seul département du Cantal. Le risque relatif comme le score associés diminuent progressivement au fil du temps tout en restant très élevés tout au long de cette période. En semaine 50 on voit apparaitre un nouveau cluster couvrant l'Île-de-France ainsi que la Haute-Normandie, l'Eure-et-Loir, l'Oise et la Somme. La semaine suivante il recouvre la moitié Ouest du pays, avant de s'étendre à toute la métropole au cours de quatre semaines consécutives. Il régresse ensuite vers l'Ouest en semaines 56 et 57. De la semaine 58 à la semaine 102 on observe un cluster correspondant au seul département des Pyrénées Orientales, avant de retrouver pour les deux dernières semaines de l'analyse le cluster du Cantal. Dans ces deux derniers cas le risque relatif comme le score associés sont importants mais pour le premier, il est en décroissance progressive, tandis que le second reste globalement stable.

Le tableau de résultats détaillé est consultable en annexe.

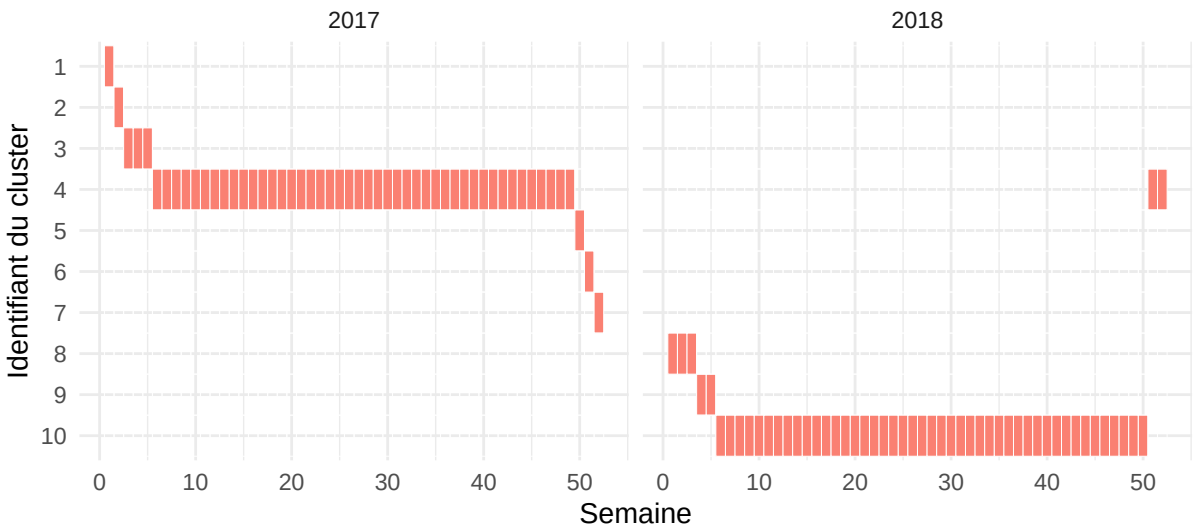

FIG. 9 : Cluster significatif le plus probable à chaque temps de l'analyse, se reporter à la figure suivante pour visualiser la localisation du cluster correspondant

5.1.3 Discussion

L'épidémie de grippe de l'hiver 2016-17 est détectée dès la première semaine de suivi, ce qui est cohérent avec les données de niveau d'alerte de l'époque. En effet, cette épidémie a commencé au cours de la semaine 49 de l'année 2016 et s'est rapidement généralisée à l'ensemble du pays. Elle n'a commencé à régresser qu'au cours de la 6e semaine de 2017, par la Bretagne, et dès la semaine suivante la seule région encore touchée était

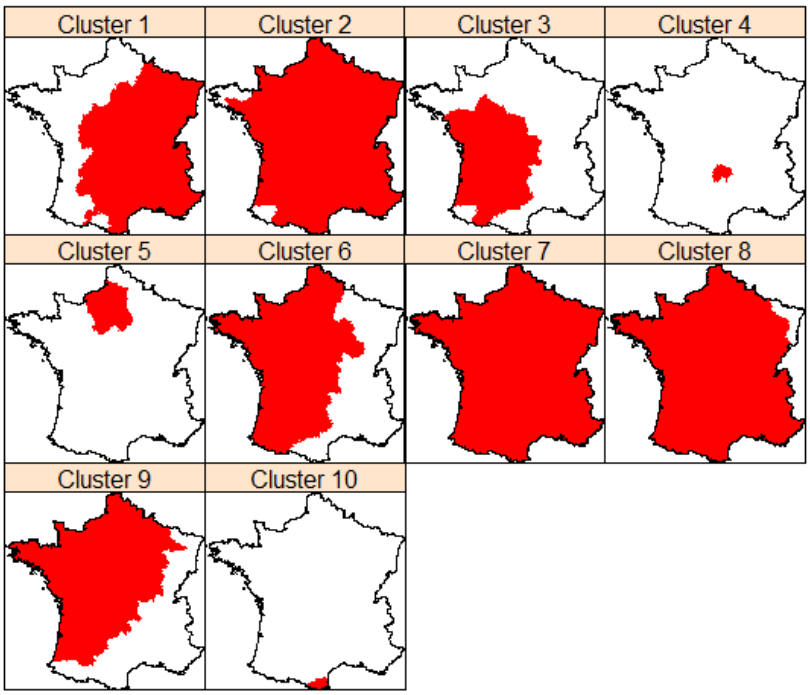

FIG. 10 : Cartes des différents clusters de consultation pour grippe détectés

les Hauts-de-France. Les clusters détectés successivement au cours des cinq premiers temps de l'analyse couvrent donc presque toute la durée de l'épidémie en 2017, mais pas forcément l'ensemble de son aire géographique. Toutefois, seuls les clusters les plus probables ont été étudiés à chaque temps t, et il est possible que la méthode ait détecté l'épidémie sous la forme de plusieurs clusters distincts plutôt que comme un seul couvrant l'ensemble du pays.

L'épidémie de grippe de l'hiver 2017-18 est détectée à partir de la semaine 50, en Île-de-France, avant une généralisation dès la semaine suivante. Elle se poursuit alors jusqu'à la 57e semaine d'analyse, soit la 5e semaine de 2018. Les données d'alerte de Santé Publique France indiquent que cette épidémie aurait débuté au cours de la 49e semaine de 2017 en Île-de-France, avant de se généraliser progressivement à l'ensemble de la métropole sur deux semaines. Elle s'est ensuite poursuivie jusqu'à la 12e semaine de 2018. L'épidémie est donc détectée relativement rapidement et avec une localisation correcte, mais ne couvre pas l'ensemble de la durée de l'épidémie. Cela pourrait s'expliquer par le passage du statut de cluster le plus probable à celui de cluster secondaire sur la fin de la période, ces derniers n'ayant pas été étudiés.

Aucune épidémie de grippe n'est détectée au cours de l'hiver 2018-19, ce qui est attendu dans la mesure où celle-ci n'a débuté qu'en 2019.

Entre ces différentes épidémies, sont identifiés deux clusters constitués de départements isolés, et correspondant manifestement à des faux positifs. On remarque qu'ils sont tous deux associés à des risques relatifs et des scores élevés, mais ayant tendance à décroitre entre deux épidémies. Ce comportement peut s'expliquer par une surincidence très importante au cours de la période épidémique. En effet, de par la nature cumulative des statistiques de scan prospectives, lorsqu'une unité spatiale comporte un excès de cas très important pendant une période donnée, pour que la statistique de test redescende sous le seuil de significativité il faut que cette période soit compensée par une période de retour à la normale voire de sous-incidence afin de diluer le risque relatif observé dans le cluster. Mais dans notre situation, le phénomène observé étant cyclique, le nombre de cas attendu devient très faible en période non-épidémique, et la contribution de ces périodes au risque relatif global dans le cluster est également diminuée. Il en résulte une décrue plus lente du score après la période épidémique qui, dans les cas extrêmes, empêche le passage sous le seuil de significativité. Une méthode d'évitement de ce phénomène aurait pu être d'imposer une limite maximale à la fenêtre temporelle utilisée, toutefois dans ce cas des faux positifs persistent tant que l'épidémie n'est pas complètement sortie de la fenêtre de balayage.

Contrairement à ce qui est réalisé dans le cadre de la surveillance temporelle des épidémies de grippe, le nombre de cas attendus a ici été estimé sans exclure les périodes épidémiques ou les taux les plus élevés observés sur la période d'apprentissage (52). Il en résulte un biais positif dans les estimations ayant pu entrainer une perte de puissance, pouvant expliquer en partie la faible capacité à couvrir l'ensemble de la période épidémique lors des grippes saisonnières. Toutefois, cette méthode a aussi l'avantage de réduire le risque de faux positif associé à l'incertitude sur les estimations des attendus, comme l'a montré Neill (33).

Ce dernier a proposé différentes méthodes permettant de compenser l'absence de prise en compte de cette incertitude dans les méthodes de détection de cluster : utiliser un estimateur biaisé positivement, fixer un seuil de significativité plus bas (en termes de risque alpha) ou utiliser des scores obtenus sur données historiques pour fixer le seuil de significativité (33,53). La méthode qu'il privilégie est la dernière citée, mais elle nécessite une profondeur historique sans épidémie très importante qui n'était pas disponible ici. La deuxième méthode proposée nécessite elle une multiplication par dix du nombre de simulations de Monte Carlo réalisées à la phase d'inférence pour chaque gain d'une décimale en précision pour la valeur-p et, au vu des scores obtenus par les faux positifs, un seuil particulièrement bas aurait été nécessaire. Par ailleurs, ces deux méthodes ne permettent pas de prendre en compte les situations de persistance de score élevé pour les phénomènes cycliques observées ici, car elles ne modifient que le seuil de significativité tandis que sans pénalité le score continuera à augmenter à chaque épidémie. Enfin, il préférait éviter l'utilisation d'estimateur biaisé en raison de la perte de puissance associée, toutefois il n'avait évalué que des estimateurs basés sur le maximum des valeurs antérieures, ce qui constitue un biais supérieur à celui utilisé ici.

En conclusion, l'analyse des données de consultation aux urgences pour grippe permettait effectivement de détecter les deux épidémies hivernales survenant sur la période étudiée. Mais si cette détection était relativement rapide, le cluster correspondant disparaissait trop précocément par rapport aux données de référence. Il était alors supplanté par des faux positifs qui entrainaient un signal significatif sur l'ensemble de la période étudiée, malgré une méthode d'estimation des attendus plutôt conservatrice.

5.2 Maladie à coronavirus de 2019 (Covid-19)

La maladie à coronavirus de 2019 (Covid-19) est une pathologie infectieuse émergente découverte en décembre 2019 et causant des pneumonies (54–56). Elle est provoquée par le virus SARS-Cov-2 (*Severe Acute Respiratory Syndrom Coronavirus* 2), virus à ARN (Acide Ribonucléique) simple brin de la famille des *Coronaviridae*, et serait apparenté au coronavirus de chauve-souris RaTG13 (56). Ses principaux symptômes, observés chez les patients hospitalisés en Chine, sont la fièvre (88,7 %) et la toux (67,8 %), pouvant s'accompagner plus rarement de troubles digestifs tels que nausées et vomissements (54). Une part des cas difficile à estimer serait asymptomatique. A partir de la population isolée du paquebot *Diamond Princess*, touché par l'épidémie de Covid-19, le taux d'hospitalisation parmi les personnes infectées a été estimé à 2,6 %, et le taux de mortalité à 0,53 % (57). Les principaux facteurs de risque identifiés de formes graves et de décès sont l'âge et la présence de comorbidités associées (54,57).

L'épidémie est apparue dans la région de Wuhan (Hubei) en Chine en novembre ou décembre 2019 (55), avant de se propager en Europe à partir de la fin du mois de février 2020, notamment en Italie, France et Espagne. Toutefois, la circulation locale du virus en France pourrait avoir commencé dès le mois de janvier (58). La stratégie de contrôle adoptée par la plupart des pays fortement touchés par la pandémie a reposé sur la mise en place de mesures de distanciation sociale et notamment de confinement de l'ensemble de

la population. En France, la décision de fermeture des lieux publics non indispensables est entrée en vigueur le 15 mars 2020, puis le confinement a été instauré à partir du 17 mars. A partir du 11 mai, la France entamera une phase de déconfinement progressif au cours de laquelle les écoles rouvriront de façon échelonnée dans le temps et avec des effectifs réduits, et impliquant également une réouverture progressive des commerces. Toutefois, les premières estimations indiquent qu'au 11 mai seule 5,7 % de la population française aura été immunisée, et que le virus sera encore en situation de circulation active (57). C'est pourquoi le gouvernement français a décidé de mettre en place le déconfinement de façon différenciée en fonction des territoires et de maintenir une surveillance de l'épidémie afin de pouvoir rétablir des mesures plus restrictives si nécessaire. La distinction des territoires reposera sur le nombre de nouveaux cas détectés au cours des 7 jours précédents, le taux d'occupation des unités de réanimation et de soins intensifs, et le nombre de tests diagnostics réalisés.

L'objectif de ce travail était de pouvoir identifier des clusters de surincidence de Covid-19 afin d'éclairer la décision publique dans un contexte de déconfinement. L'incidence de la maladie était évaluée via un proxy : l'hospitalisation en réanimation pour Covid-19.

5.2.1 Matériel et méthodes

Les données étaient issues de Santé Publique France, qui les publie quotidiennement depuis le 19 mars 2020 sur la plate-forme gouvernementale d'accès aux données ouvertes (data.gouv). Elles contiennent exclusivement des données hospitalières concernant les hospitalisations, les passages en réanimation, les décès et les sorties de patients Covid-19, et sont agrégées au département. Les nombres de cas incidents dans chaque catégorie sus-citée sont fournis pour chaque jour à partir de la date du 19 mars, soit deux jours après le début du confinement.

Pour cette analyse, les données ont été figées au 26 avril. Pour chaque jour un taux d'incidence était calculé pour 100000 habitants à partir des populations départementales, afin de produire des cartes d'incidence pour chaque date étudiée. Les couleurs des départements étaient déterminées par quintiles d'incidence quotidienne observée sur l'ensemble de la période, les deux premiers étant regroupés car plus de 20 % des incidences observées étaient nulles. Aucune standardisation n'était réalisée.

S'agissant de détecter des clusters pour une pathologie émergente, il n'était pas possible de calculer des effectifs attendus à partir de données historiques, l'analyse a donc été réalisée avec la méthode de statistique de scan conditionnelle pour distribution de Poisson de Kulldorff (32). La fenêtre de scan utilisée était circulaire avec une taille limitée de façon à ne pas contenir plus de 50 % de la population. Les analyses débutaient le 19 mars 2020 et se poursuivaient quotidiennement jusqu'au 26 avril en imposant une fenêtre temporelle maximale de 7 jours. Cette durée correspondait à la fenêtre utilisée par le gouvernement français pour comptabiliser les cas afin de fixer le niveau épidémique de chaque département, déterminant les mesures de distanciation sociale à appliquer dans le cadre du déconfinement. Des clusters secondaires étaient recherchés selon le même principe que pour le cluster le plus probable, c'est-à-dire que la valeur-p associée était calculée à partir

de la distribution du score maximal observé dans les simulations de Monte Carlo, comme recommendé par Kulldorff (32).

Ces analyses ayant vocation à être répétées quotidiennement tout au long du suivi de l'épidémie et donc possiblement jusqu'à obtention d'un vaccin, la durée prévisible de la surveillance est d'au moins 12 à 18 mois. Le risque de faux positifs associé à ces tests multiples a été contrôlé par fixation d'un seuil de significativité plus strict via la méthode de Sidak (59). Il s'agissait de garantir un risque de première espèce de 5 % sur une durée d'un an, soit 365 jours, et le seuil de significativité était donné par :

$$\bar{\alpha} = 1 - (1 - \alpha)^{\frac{1}{m}} = 1 - (1 - 0.05)^{\frac{1}{365}} \approx 0.00014$$

avec $\bar{\alpha}$ le risque de première espèce pour chaque analyse, α le risque global que l'on souhaite garantir et m le nombre d'analyses. En conséquence, les analyses ont été réalisées avec 9999 simulations de Monte Carlo, nombre minimal nécessaire pour atteindre une précision suffisante.

Les clusters détectés précédemment étaient ensuite représentés sur des cartes.

Les analyses statistiques ont été réalisées avec le logiciel R version 3.6.1 (2019-07-05) et le package scanstatistics version 1.0.1 (49).

5.2.2 Résultats

Seules quelques cartes d'incidences sont représentées. Elles montrent une augmentation de l'incidence des hospitalisations en réanimation au cours des 15 premiers jours, puis une régression progressive à partir de la troisième semaine de confinement. Les principales zones touchées sont l'Est ainsi que l'Île-de-France et la Picardie, et dans une moindre mesure la Côte méditerranéenne. L'augmentation initiale touche l'ensemble du pays, même si l'Ouest et le Centre restent plus épargnés.

Les clusters détectés couvrent initialement l'ensemble du quart Nord-Est de la métropole, s'étendant par moments jusqu'à la région lyonnaise, qui apparait en tant que cluster secondaire à certains temps de l'analyse. A partir du 23e jour de suivi, le cluster du quart Nord-Est se scinde en deux : un cluster primaire couvrant la majeure partie de l'Île-de-France, et un cluster secondaire regroupant l'Alsace et la Lorraine, avant de régresser progressivement pour ne plus couvrir que le Haut-Rhin et de disparaitre à t28 et t29. En Île-de-France, du 28e au 31e jour, le cluster significatif ne couvrait plus que les quatre départements centraux, ainsi que le Val d'Oise à t28. A partir du 30e jour de suivi, un nouveau cluster secondaire apparait en Alsace-Lorraine, puis s'étend progressivement jusqu'à couvrir à nouveau le tiers Nord-Est du pays à t35, du Nord à la région lyonnaise et l'Île-de-France. Au cours des jours suivants il régresse légèrement au sud. Lors des trois derniers jours du suivi, à t38 et t39, un cluster secondaire apparait dans les Bouches-du-Rhône.

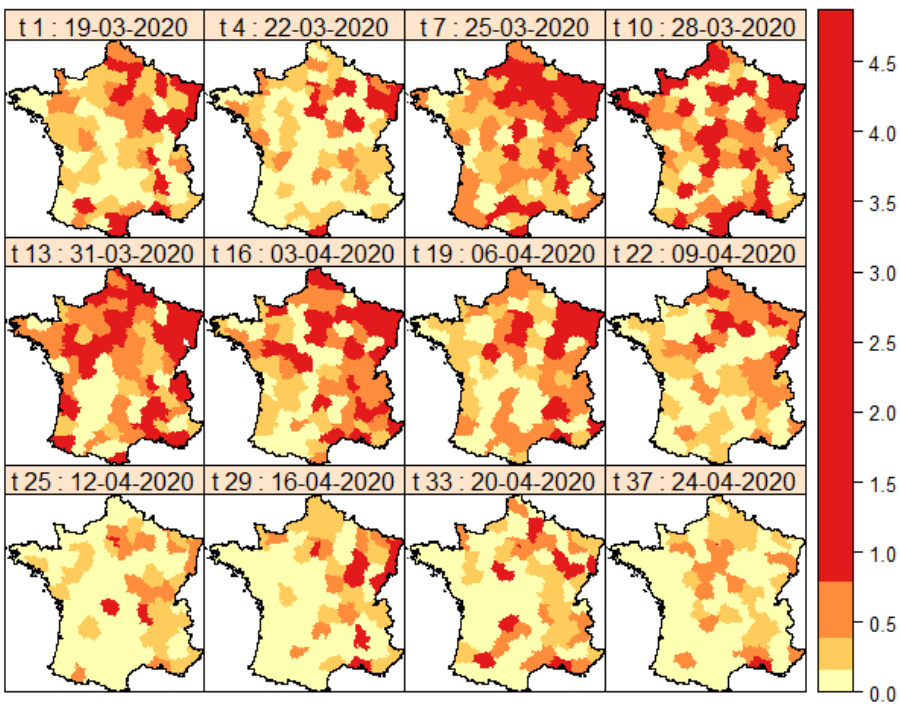

FIG. 11 : Cartes d'incidence (pour 100000 habitants) de l'hospitalisation en réanimation pour Covid-19 à différents temps

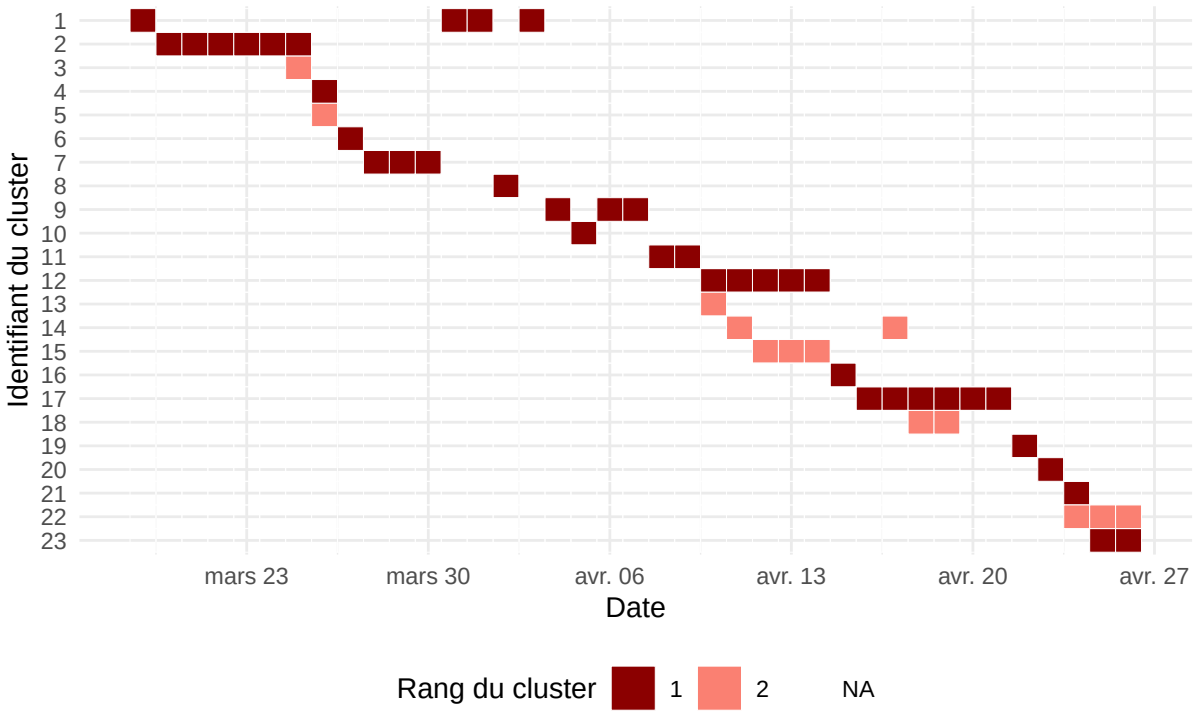

FIG. 12 : Clusters significatifs à chaque temps de l'analyse du Covid-19, se reporter à la figure suivante pour visualiser la localisation du cluster correspondant

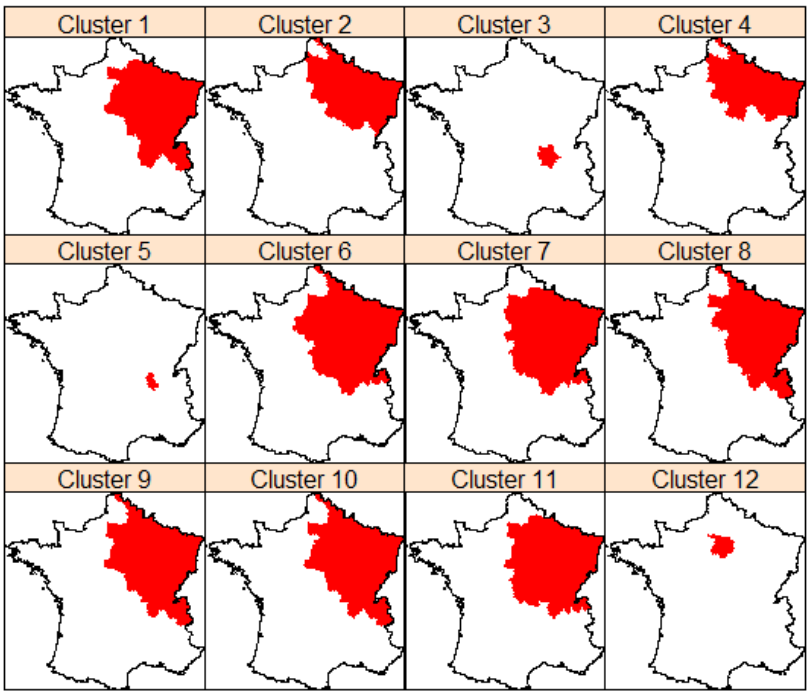

FIG. 13 : Cartes des 12 premiers clusters de Covid-19 détectés (t1 à t22)

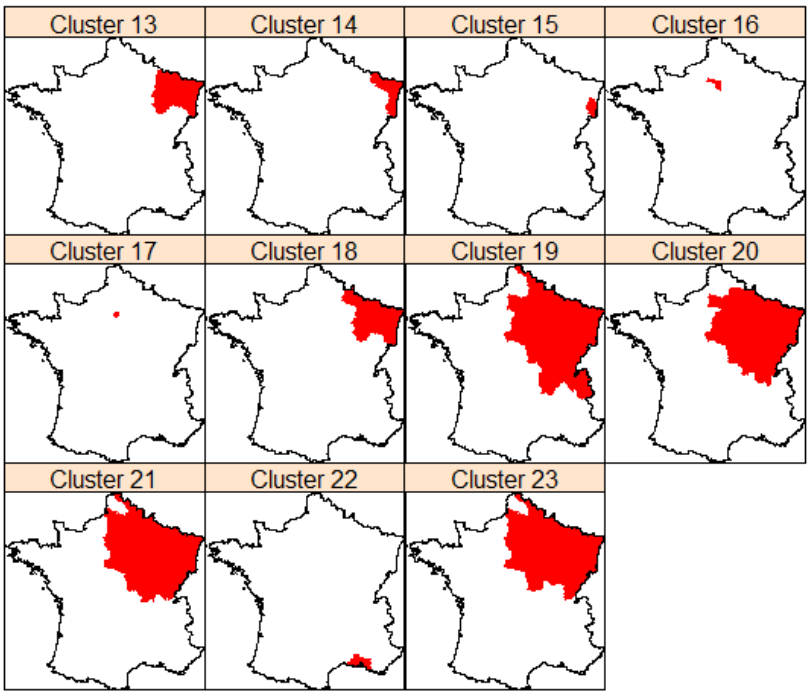

FIG. 14 : Cartes des clusters de Covid-19 détectés n°13 à 23 (t23 à t39)

5.2.3 Discussion

Les clusters détectés au cours des 21 premiers jours du suivi couvraient les régions Grand-Est, Île-de-France, la Picardie, la Bourgogne-France-Comté et parfois le nord de l'Auvergne-Rhône-Alpes. Ces territoires ont effectivement abrité les trois principaux clusters de transmission de la Covid-19 identifiés en France à la fin du mois de février et au début du mois de mars (60). Ceux-ci étaient localisés dans le sud de l'Oise, dans les environs de Mulhouse (Haut-Rhin) et en Haute-Savoie. Les régions Grand-Est et Île-de-France étaient les plus touchées au cours de l'épidémie, suivies des Hauts-de-France, de la Bourgogne-Franche-Comté et d'Auvergne-Rhône-Alpes (60). Un cluster de transmission avait également été détecté dans le Morbihan au début du mois de mars, mais ensuite la Bretagne n'est jamais apparue comme une zone de diffusion importante de l'épidémie. Les actions d'identification et d'isolement des cas-contacts ont peut-être permis de limiter sa propagation sur le territoire breton.

Du 23e au 30e jour du suivi, la taille des clusters régresse et seule la métropole de Paris persiste pendant quelques jours. Ce constat est cohérent avec la régression de l'épidémie observée par Santé Publique France après trois semaines de confinement (initié le 17 mars, soit deux jours avant le début des analyses) (60). Le confinement a peut-être contribué à homogénéiser à la baisse l'incidence de la maladie à l'échelle nationale. Le délai de trois semaines observé est dû à la durée d'incubation de 3 à 14 jours et au délai de développement de la forme sévère. Finalement, à partir du 31e jour et jusqu'à la fin du suivi, un cluster ré-émerge dans le quart Nord-Est de la métropole. Il est le témoin, dans une phase de plateau où le nombre de nouveaux cas poursuit sa baisse mais de façon plus lente, de la persistance d'un risque plus élevé dans les régions ayant abrité une chaîne de contamination. Le confinement n'ayant pas été total, comme dans le Hubei en Chine, la propagation du virus n'a pas été complètement interrompue dans ces territoires. Ce cluster n'indique pas une ré-émergence, le nombre de cas poursuivant sa décrue ou *a minima* sa stabilisation, mais plutôt un risque de seconde vague de transmission.

L'indicateur adopté pour le suivi de l'épidémie de Covid-19, l'hospitalisation en réanimation, présente quelques défauts. Il s'agit d'un indicateur tardif, la durée d'incubation de la maladie étant de l'ordre d'une à deux semaines, délai auquel il faut ajouter une à deux semaines supplémentaires pour que l'aggravation des cas entraine leur transfert en réanimation. Par ailleurs, le nombre de lits d'hospitalisation en soins intensifs étant limité, lorsque les capacités de prise en charge des établissements sont atteintes ou dépassées, l'indicateur est faussé avec une sous-estimation dans les territoires les plus tendus. En revanche, compte tenu de la standardisation des procédures en soins intensifs, il est probable qu'il n'y ait pas eu de différence majeure dans la probabilité de transfert en réanimation en cas de forme sévère, en-dehors des situations de saturation des services concernés. De même, il est probable que les patients graves aient eu la même propension à contacter les services de secours quel que soit leur territoire de résidence, ce qui est moins évident pour les patients peu symptomatiques.

D'autres indicateurs auraient pu être utilisés. Les décès de patients Covid-19 constituent un indicateur moins sujet aux biais associés au manque de capacités hospitalières, mais

qui a le défaut d'être encore plus tardif que le transfert en réanimation. L'hospitalisation pour Covid-19 pourrait permettre de détecter plus précocément le développement d'un nouveau cluster, mais la décision d'hospitalisation étant moins standardisée que le transfert en soins intensifs, elle est davantage susceptible d'être impactée par le niveau de saturation des établissements de soins. A symptômes identiques, dans les territoires les plus touchés un malade aura moins de chances d'être hospitalisé que dans un territoire relativement épargné, ce qui tend à réduire l'écart observé entre leurs taux d'hospitalisation respectifs. Par ailleurs, tous ces indicateurs ne concernant que des patients hospitalisés, ils excluent par définition les personnes n'ayant pu être dépistées et décédées hors d'un établissement de soin. Le dernier indicateur envisageable est le nombre de patients diagnostiqués. Toutefois, celui-ci est dépendant du nombre de test réalisés, or en France la stratégie a rapidement consisté à cibler les dépistages sur les patients les plus gravement atteints, donc hospitalisés. La stratégie de limitation des dépistages a probablement conduit à une forte hétérogénéité spatiale dans la possibilité d'accéder à un test, ce qui entrainerait un biais dans la détection de cluster. Il est aussi à noter que la ville de Marseille a bénéficié d'un accès plus large au dépistage, du fait de décisions locales au sein de l'Institut Hospitalo-universitaire Méditerranée Infection, ce qui pourrait entrainer un biais de sur-estimation des cas positifs dans cette localité.

Il a été choisi de fixer un maximum à la fenêtre de scan temporelle car l'objectif était ici d'identifier les territoires les plus à risque au moment de l'analyse, ou de détecter une réémergence. L'inclusion de l'ensemble des données depuis le début du suivi aurait accordé un poids important aux régions fortement touchées lors de la première vague épidémique, ce qui aurait pu masquer l'apparition récente d'un cluster à risque faiblement augmenté. Le délai de 7 jours correspondait à celui choisi par le gouvernement français pour la quantification du risque à l'échelle départementale dans le cadre du déconfinement. Un délai plus court limiterait la capacité à détecter des clusters de risque faible à modéré, qui nécessitent généralement un cumul des cas dans le temps pour atteindre la significativité.

Aucune standardisation n'a été effectuée alors que la Covid-19 est associée à l'âge, et que le sexe est associé aux comorbidités pouvant impacter l'évolution de la maladie. Il pourrait en résulter des variations territoriales d'incidence de transferts en réanimation qui ne seraient pas dues à une circulation plus importante du virus, mais à une plus grande fragilité de la population exposée. Toutefois les résultats obtenus ne présentaient pas d'incohérence quant à la localisation des clusters détectés. Par ailleurs, les données fournies par Santé Publique France ne permettaient pas d'obtenir des effectifs à la fois par sexe et classe d'âge. De plus, les contrôles d'incohérence fournis faisaient état de 12 à 39 % d'effectifs départementaux quotidiens non concordants avec le total des cas par sexe, selon l'indicateur choisi. Compte tenu du faible nombre de patients observés à l'échelle quotidienne dans chaque département, ces erreurs dans la base de données auraient pu conduire à une mauvaise estimation du nombre de cas attendus par standardisation. De tels niveaux d'incohérence sont compréhensibles dans le cadre d'un système de surveillance exhaustif à l'échelle nationale et mis en place en urgence pour une pathologie inconnue quelques mois auparavant. Par ailleurs les mises à jour quotidiennes compliquent la mise en place d'un contrôle qualité efficace.

La prise en compte de la répétition des analyses dans le temps conduisait à un seuil de significativité particulièrement bas (0.00014). Il a pu en résulter une perte de puissance pour détecter des clusters. Néanmoins, tous les territoires dans lesquels des chaînes de contamination avaient été identifiées précédemment ont correctement été détectés par la méthode. Par ailleurs, un seuil de significativité de ce niveau imposait d'obtenir une précision d'au moins 4 décimales dans l'estimation de valeur-p. Pour cela, il a fallu porter à 9999 le nombre de simulations de Monte Carlo réalisées pour chaque analyse. Obtenir une précision de 5 décimales aurait nécessité 99999 simulations de Monte Carlo, soit une multiplication du temps de calcul par 10. Ce dernier peut alors rapidement devenir un frein à la mise en oeuvre d'un système de surveillance intégrant des méthodes de détection de cluster, en particulier dans une situation épidémique où les résultats des analyses sont attendus quotidiennement par les médias et le grand public. En particulier, le temps de calcul associé à ces méthodes étant dépendant du carré du nombre d'unités spatiales du territoire étudié, le maintien d'un tel niveau de précision dans l'estimation de la valeur-p serait difficilement compatible avec une augmentation de la précision spatiale, qui impliquerait un accroissement important du nombre d'unités spatiales.

En conclusion, la méthode de statistique de scan conditionnelle pour distribution de Poisson a permis de détecter des clusters d'incidence de transfert en réanimation pour Covid-19 pertinents dans le cadre du suivi de l'épidémie. Les méthodes de détection de cluster pourraient apporter un complément utile aux méthodes de la surveillance épidémiologique temporelle afin d'éclairer la décision publique tout au long de la mise en place du déconfinement progressif de la France.

5.3 Mortalité toutes causes

La mortalité est l'un des principaux indicateurs suivis en routine pour surveiller l'état de santé des populations. Le suivi annuel à l'échelle nationale et l'étude des inégalités territoriales permet d'évaluer l'impact et les évolutions de long terme des pathologies chroniques, mais le suivi à court terme (hebdomadaire voire quotidien) a également un intérêt majeur dans le suivi des événements sanitaires exceptionnels. En effet, lorsqu'une crise sanitaire apparait, il est primordial d'évaluer rapidement la menace qu'elle représente afin de prendre des décisions de santé publique adaptées et proportionnées (4). La mortalité associée à ces événements est l'un des principaux indicateurs de leur sévérité et donc un élément important dans la décision publique (61,62). Par exemple, parmi les deux pandémies qui ont touchées la France depuis 15 ans, la grippe A/H1N1 n'a pas entrainé de surmortalité significative (62), alors que la maladie à coronavirus de 2019 (Covid-19) a provoqué au moins 16653 décès en France du mois de février au 14 avril 2020 (60).

En France, l'impact sanitaire de la canicule de 2003 et le retard pris dans la gestion de crise à l'époque ont largement concouru à la mise en place d'un système national de surveillance syndromique (SurSaUD) dédié à l'identification de situations sanitaires exceptionnelles de façon aspécifique (12,61). Il se base sur les données de consultation aux urgences de services volontaires (réseau OSCOUR) ainsi que les consultations de SOS médecins en ambulatoire et les données de mortalité toutes causes. Toutefois ce système

de surveillance repose essentiellement sur des méthodes de suivi temporel par méthode de Shewhart à l'échelle nationale et des régions (61). Or, comme évoqué précédemment, il a été montré par Neill que la réalisation d'une surveillance séparément dans plusieurs unités spatiales entrainait une augmentation du délai de détection des événements inhabituels. Il pourrait donc être intéressant d'intégrer ce type de méthodes au système de surveillance.

L'objectif de ce travail était d'évaluer la possibilité de mettre en oeuvre une méthode de détection de cluster prospective dans le cadre de la recherche d'épisodes de surmortalité inhabituels.

5.3.1 Matériel et méthodes

Les données étaient fournies par l'Institut National des Statistiques et des Etudes Economiques (Insee), sur la plate-forme gouvernementale de diffusion des données ouvertes (data.gouv), en vertu de l'avis de la Commission d'Accès aux Documents Administratifs (CADA) du 17 mai 2019. Il s'agit de l'intégralité des actes de décès enregistrés par l'Insee depuis 1970. Les fichiers contiennent l'identité de la personne décédée, son sexe, ses dates et lieux de naissance et de décès, mais pas sa cause médicale. Toutes les données ont été anonymisées pour l'analyse.

Les nombres de décès observés étaient agrégés au niveau départemental et à la semaine de survenue. L'objectif étant l'identification d'épisodes de surmortalité inhabituels, le calcul des effectifs attendus par standardisation à l'échelle nationale a été exclu. Ils ont donc été obtenus dans chaque département par une régression périodique de Poisson incluant une tendance linéaire et une périodicité annuelle, selon les recommandations du consortium européen dédié au suivi de la mortalité Euro-MOMO (*European monitoring of excess mortality for public health action*) (62). Le modèle était donc le suivant :

$$\ln(Y_t) = a_0 + a_1 t + \gamma \cos(2\pi t/n) + \delta \sin(2\pi t/n) + \epsilon_t$$

avec Y_t la variable à expliquer (le nombre de décès), a_0 l'intercept, a_1, γ et δ les coefficients associés aux variables explicatives, t le temps en semaines, n la durée d'une période (ici 52,179 pour une année) et ϵ_t l'erreur résiduelle.

Le modèle était entrainé sur une période historique de 5 ans de 2014 à 2018, puis les paramètres obtenus étaient utilisés pour prédire le nombre de cas attendus sur la période d'intérêt de l'étude, de janvier 2019 à mars 2020.

Une analyse de détection de cluster prospective était réalisée avec la méthode non conditionnelle pour distribution de Poisson (33) et une fenêtre de balayage circulaire, selon la méthode de Kulldorff (32). Aucun seuil de population ou de taille maximale n'a été imposé. En effet, étant donné que la méthode de Neill ne compare pas l'intérieur et l'extérieur du cluster, détecter un cluster à un endroit n'équivaut pas à détecter un cluster de risque diminué ailleurs. Ces analyses hebdomadaires portaient sur la période de janvier 2019 à mars 2020, chacune prenant en compte l'ensemble de la période antérieure depuis le

début des analyses. Les données de la dernière semaine de mars 2020 ont été retirées de l'analyse car, compte tenu du délai de déclaration, elles comportaient trop de décès manquants.

Le risque de faux positif associé aux situations de tests multiples a été contrôlé par la méthode de Kulldorff (32,42) de façon à garantir un risque de première espèce de 5 % sur 52 analyses, soit une année de suivi.

Des cartes de rapports de mortalité ont été produites pour chaque date étudiée (en 2019-20) à partir des effectifs observés et attendus calculés précédemment, et d'autres cartes afin de visualiser les clusters détectés par la méthode de statistique de scan.

Les analyses statistiques ont été réalisées avec le logiciel R version 3.6.1 (2019-07-05) et le package scanstatistics version 1.0.1 (49).

5.3.2 Résultats

Des cartes de rapports de mortalité sont présentées à différentes dates de la période étudiée. Elles montrent clairement des périodes de mortalité importante en hiver et en été, tandis que celles du printemps et de l'automne sont plus proches de ce qui est attendu. La dernière carte présentée correspond à la dernière semaine étudiée, la 3e de mars 2020, pendant la pandémie de Covid-19. Elle montre une mortalité globalement plus faible que ce qui était attendu à l'échelle nationale, mais avec quelques foyers de surmortalité importante mais localisés : en Alsace-Lorraine, dans le Nord-Ouest de l'Île-de-France et le Sud de la Picardie.

Les clusters détectés correspondaient à 3 périodes distinctes :

- du 7 janvier au 21 avril 2019 (t2 à t16) : épidémie de grippe saisonnière (du 7 janvier au 3 mars)
- du 24 juin au 1 septembre 2019 (t26 à t35) :
  - canicule européenne de juin 2019 (du 23 juin au 8 juillet)
  - canicule européenne de juillet 2019 (du 21 juillet au 1er août)
- du 9 au 22 mars 2020 (t63 à t64) : pandémie de Covid-19

Concernant la localisation des clusters identifiés, le premier d'entre eux apparait dès la 2e semaine de 2019 sur la majeure partie du pays, avant de le recouvrir presque entièrement au cours des semaines suivantes, puis de régresser par le Nord et de se concentrer sur la région Auvergne-Rhône-Alpes dans ses dernières semaines. Le second, apparait au centre de la métropole, avant de s'étendre à la majeure partie du pays à l'exception de sa frange Nord. Puis, après avoir régressé pendant une semaine, il se déplace sur l'ensemble de la moitié Nord du pays à la fin du mois de juillet, pour ensuite se rejoindre à l'Est pour les semaines suivantes. Le dernier cluster est lui très localisé sur le département du Haut-Rhin à partir de la deuxième semaine de mars 2020.

Le tableau de résultats détaillés est consultable en annexe.

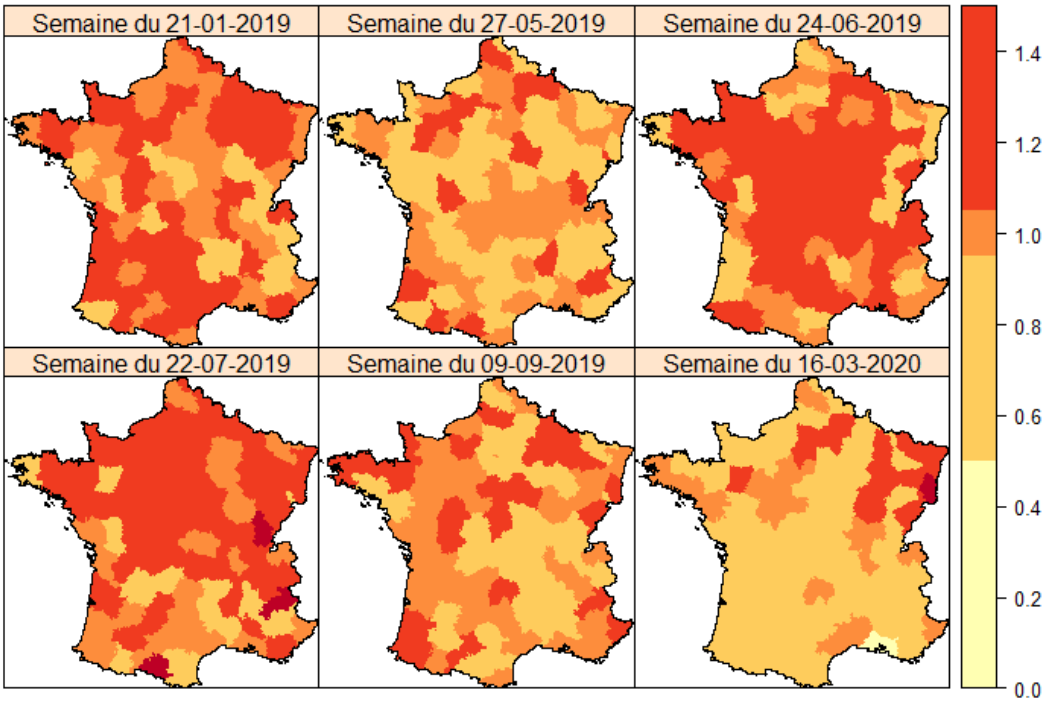

FIG. 15 : Cartes de rapports de mortalité à différents temps

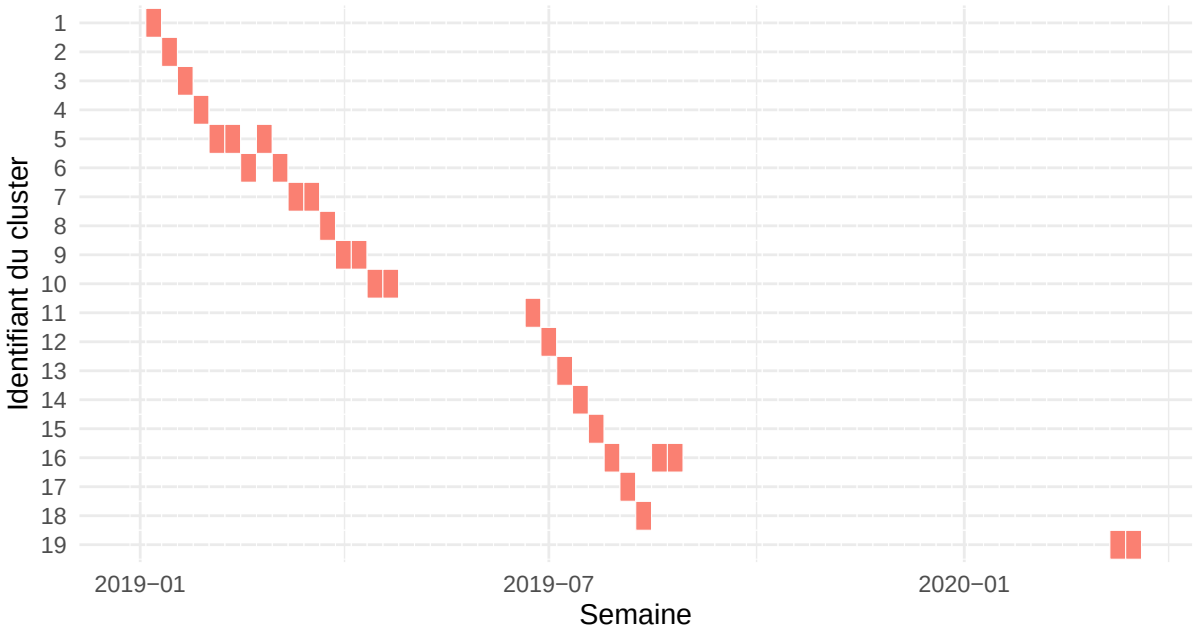

FIG. 16 : Cluster significatif le plus probable à chaque temps de l'analyse, se reporter à la figure suivante pour visualiser la localisation du cluster correspondant

5.3.3 Discussion

Tous les clusters identifiés correspondent bien à des épisodes de surmortalité connus : l'épidémie de grippe saisonnière de l'hiver 2018-19, les deux canicules successives de juin-juillet 2019 et la pandémie de Covid-19 en mars 2020. Aucune surmortalité n'est identifiée au cours de l'hiver 2019-20, ce qui est en ligne avec les observations de Santé Publique France jusqu'en mars (63).

Concernant la localisation des clusters détectés, celui du début d'année couvre rapidement la totalité du pays jusqu'à la semaine 8 avant de régresser par le Nord, puis de persister autour de la région Auvergne-Rhône-Alpes des semaines 13 à 16. L'épidémie de grippe saisonnière de l'hiver 2018-2019 a débuté en Occitanie pour couvrir progressivement toute la métropole au cours des trois premières semaines de l'année (50). Elle a ensuite persisté jusqu'à la semaine 9 avant de régresser par le Nord en semaine 10. La localisation des clusters identifiés en début d'année est donc cohérente avec les données de suivi de la grippe pour la saison correspondante, mais ils tendent à persister au-delà de la fin de l'épidémie.

Les clusters correspondant la canicule de juin 2019 s'étendent sur le centre-est du pays lors de la 26e semaine, avant de couvrir pendant deux semaines les deux-tiers sud de la métropole, jusqu'au bassin parisien, puis de régresser à nouveau sur le Centre-Est en semaine 29. Cette géographie correspond assez bien aux cartes d'anomalie de température pour la dernière semaine de juin, par ailleurs les alertes météorologiques déclenchées

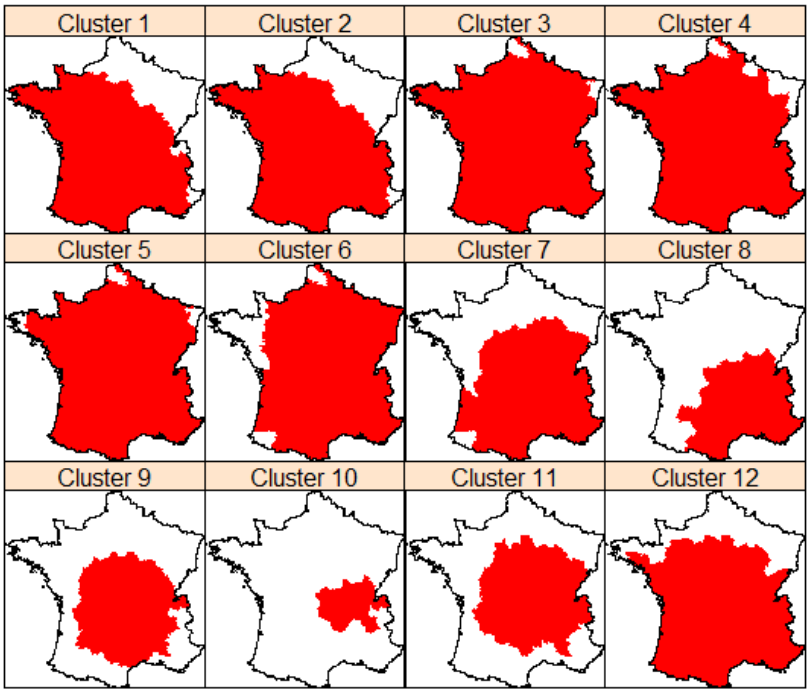

FIG. 17 : Cartes des 12 premiers clusters détectés

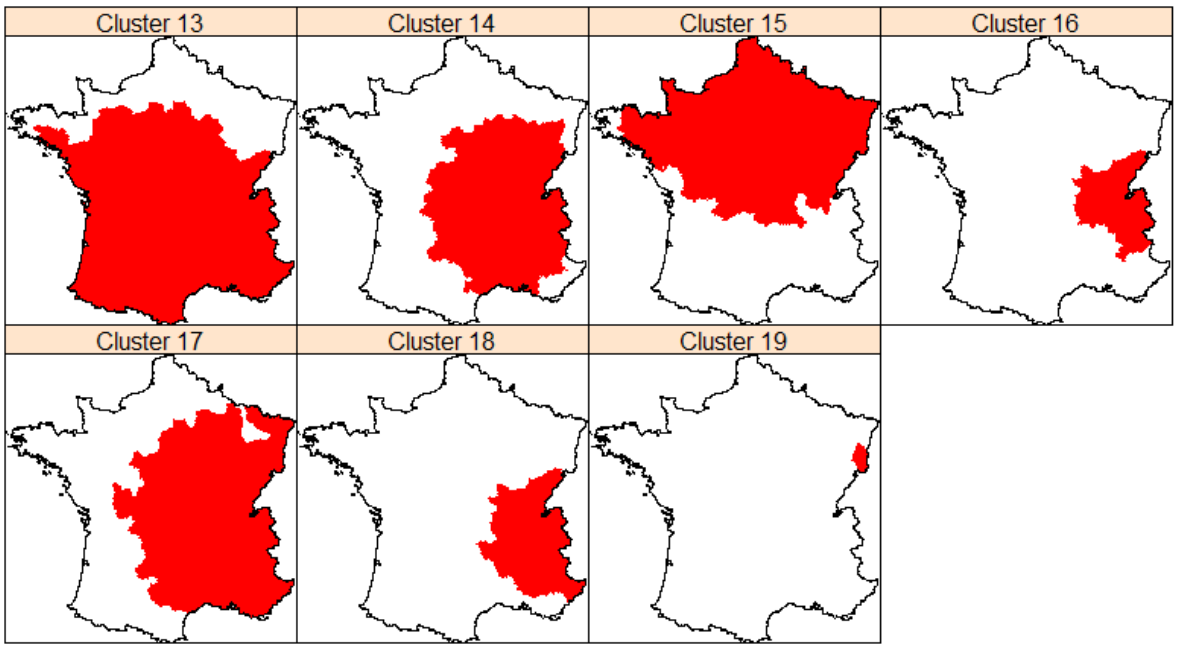

FIG. 18 : Cartes des clusters détectés n°13 à 19

pour cette canicule couvraient presque tout le pays à l'exception des régions bordant la Manche et d'une partie de l'Aquitaine. La canicule de fin juillet 2019 touche elle davantage de la moitié Nord de la France, mais seules la Bretagne et la Côte Méditerranéenne sont réellement épargnées. Cette géographie est retrouvée dans la localisation du cluster n° 15 en semaine 30. Un cluster reste ensuite significatif au mois d'août dans l'Est de la France, autour de la région Auvergne-Rhône-Alpes, ce qui pourrait s'expliquer par l'importance de la surmortalité observée dans ce secteur touché successivement par les deux épisodes caniculaires. Cette interprétation est aussi supportée par la durée du cluster qui le fait débuter au mois de juin.

Le dernier cluster détecté, celui du Haut-Rhin correspond effectivement au foyer principal de l'épidémie de Covid-19 en France, autour de la ville de Mulhouse.

On peut noter que les clusters sont détectés assez rapidement après le début de la crise sanitaire correspondante, mais qu'ils restent significatifs longtemps après leur fin. Ce comportement peut s'expliquer par la nature cumulative des statistiques utilisées et le fait que le phénomène surveillé soit cyclique, comme évoqué dans la discussion de l'analyse des consultations pour grippe. De fait, lorsqu'un cluster était très fortement significatif, même si les nouveaux effectifs de décès quotidiens reviennent à la normale, il faut un certain temps pour que la statistique de scan de la fenêtre recouvrant ce cluster redescende sous le seuil de significativité. Dans le cadre de la surveillance épidémiologique, ces méthodes sont évaluées sur leur capacité à détecter l'apparition d'un cluster, mais ne sont jamais évaluées sur leur capacité à ne pas détecter un cluster récemment disparu.

Contrairement à ce qui est proposé dans le cadre de la méthode européenne pour le suivi et l'estimation des épisodes de surmortalité (62), le nombre de décès attendus a été estimé en utilisant l'ensemble des données disponibles, sans retirer les périodes à risque de surmortalité (été et hiver). Il en résulte un biais positif dans les estimations qui peut entrainer une perte de puissance, mais aussi un meilleur contrôle du risque de faux positif. En effet, les méthodes de détection de cluster ne permettant pas de prendre en compte l'incertitude sur l'estimation des attendus, elles peuvent produire plus de faux positifs que leur risque α nominal ne le laisserait présager lorsque l'analyse porte sur des données réelles à forte variabilité (33).

En conclusion, l'utilisation de méthode de détection de cluster sur ces données permettait bien d'identifier les principaux épisodes de surmortalité connus sur la période étudiée. La détection de l'apparition de ces épisodes était rapide, mais le retour à la non significativité après leur fin pouvait prendre quelques semaines. En-dehors de cette persistance post-cluster, aucun faux positif n'a été identifié.

5.4 Grossesses non désirées

En France, on compte en moyenne 2 naissances par femme au cours de la vie parmi lesquelles 0,4 seraient non prévues, dont 0,05 non désirées. A ces grossesses on peut ajouter celles qui aboutissent à une interruption volontaire de grossesse (IVG), il y en aurait en moyenne 0,5 par femme en vie entière (64). Depuis l'entrée en vigueur de la loi

Neuwirth portant sur la légalisation de la contraception en 1972, le nombre de grossesses non prévues a diminué en France de 1,23 sur la période 1973-1977, à 0,85 sur la période 2005-2009 chez les femmes de 18 à 44 ans (64). Parallèlement le nombre d'IVG reste globalement stable depuis 20 ans à environ 220000 par an (64,65). Toutefois, ce nombre moyen de 0,85 grossesses non prévues par femme reste élevé, avec de fortes inégalités sociales et territoriales associées, et sa diminution est donc toujours un objectif prioritaire de la Stratégie nationale de santé sexuelle 2017-2030 (64), et de sa feuille de route pour 2018-2020 (66).

Pour répondre à cet objectif, la stratégie nationale s'appuie en particulier sur la formation des professionnels de santé, l'amélioration de l'accès à la contraception et au soin, ainsi que la promotion de la santé. Par ailleurs, le service sanitaire a été mis en place à partir de 2018. Il vise à envoyer des étudiants en santé mener des actions de prévention et promotion sur différents sujets sanitaires, principalement dans des écoles de leur région, la contraception faisant partie des sujets possibles. Ce programme, de même que la mise en application locale au sein de chaque Agence Régionale de Santé (ARS) de la stratégie nationale, peut être l'occasion de déployer dans les territoires des actions de prévention renforcée en fonction de leurs besoins spécifiques. L'utilisation de méthodes de détection de cluster pourrait alors avoir un intérêt dans le ciblage de territoires prioritaires, mais aussi dans l'évaluation des actions mises en oeuvre à travers un objectif de diminution progressive du score associé aux clusters détectés, avant sa disparition. Un suivi régulier dans un cadre de la surveillance épidémiologique permettrait alors à la fois d'identifier de nouveaux territoires nécessitant une action renforcée, et de vérifier l'efficacité des actions mises en oeuvre pour les ajuster le cas échéant.

L'objectif de ce travail était de pouvoir détecter des territoires ayant davantage de besoins que la moyenne en matière de prévention des grossesses non désirées et de promotion de la contraception. Ces territoires étaient identifiés à partir du taux de recours à l'IVG pour grossesse non désirée.

5.4.1 Matériel et méthodes

Les données provenaient de la base de données nationale du Programme de Médicalisation des Systèmes d'Information (PMSI) pour la période 2007-2014, fournie par l'Agence Techique de l'Information Hospitalière (ATIH). Elles étaient issues d'une étude en cours de rédaction portant sur l'analyse de l'activité d'interruption volontaire de grossesse et de ses déterminants à l'échelle nationale sur cette période. La base de données avait fait l'objet d'une autorisation explicite de la Commission Nationale Informatique et Libertés (CNIL) avant la mise en place de méthodologies de référence. Les IVG étaient identifiées par le biais d'un algorithme issu des règles de codage ATIH des séjours pour IVG. Un séjour était considéré comme une IVG lorsqu'il comportait un code d'acte CCAM (Classification Commune des Actes Médicaux) parmi JNJP001 (avortement médical) ou JNJD002 (avortement chirurgical), un diagnostic principal (DP) "avortement médical" (O04) et un diagnostic associé "Difficultés liées à une grossesse non désirée" (Z640). Les interruptions médicales de grossesse étaient exclues. Seul le premier IVG était conservé, ce qui implique que pendant les premières années de la période étudiée le nombre d'IVG

incidents est surévalué par absence d'information sur les actes réalisés avant 2007. En conséquence, les deux premières années du suivi (2007 et 2008) ont été retirées.

Les données ont ensuite été agrégées par département et par mois de chaque année afin d'obtenir le nombre de cas observés. Les effectifs attendus à ces mêmes dates étaient obtenus par une procédure de standardisation indirecte. A l'échelle nationale, des taux d'incidence spécifiques par tranche d'âge de 5 ans ont été calculés pour chaque mois de chaque année. Un taux spécifique attendu était calculé pour chaque mois et année à compter de janvier 2013 à partir de la moyenne des taux spécifiques observés sur les 3 années précédentes au même mois. Puis ces taux spécifiques nationaux étaient appliqués aux populations féminines par tranche d'âge de 5 ans des unités spatiales étudiées (les départements) et sommés afin d'obtenir les nombres de cas attendus.

Des cartes de ratio d'incidence standardisé ont été produites pour chaque date étudiée (en 2013-2014) à partir des effectifs observés et attendus calculés précédemment.

Une analyse de détection de cluster prospective a été réalisée avec la méthode non conditionnelle pour distribution de Poisson (33). La fenêtre de balayage utilisée était circulaire selon la méthode de Kulldorff (32), avec un seuil de population maximale de 50 %. Les analyses débutaient en janvier 2013 et se terminaient en décembre 2014. Chaque analyse portait sur le mois en cours ainsi que tous les mois précédents depuis le début du processus (janvier 2013).

Afin de prendre en compte le risque de faux positif inhérent aux situations de tests multiples, la méthode proposée par Kulldorff (32,42), qui consistait à prendre en compte les simulations de Monte Carlo réalisées pour les analyses précédentes lors de la phase d'inférence a été utilisée. Elle était paramétrée de façon à garantir un risque de première espèce de 5 % sur 12 analyses, soit une année de suivi.

Enfin, des cartes étaient produites pour visualiser les clusters détectés par la méthode précédente.

Les analyses statistiques ont été réalisées avec le logiciel R version 3.6.1 (2019-07-05) et le package scanstatistics version 1.0.1 (49).

5.4.2 Résultats

Les cartes de ratio d'incidence standardisé témoignent d'une importante hétérogénéité spatiale dans le recours à l'IVG. On observe à tous les temps de l'analyse un taux de recours plus important sur la côte méditerranéenne, ainsi que dans le Sud-Ouest à l'exception du Pays Basque. Au cours de la deuxième moitié de 2013 et de la première moitié de 2014, on observe une surincidence dans le bassin parisien qui semble s'estomper ensuite. Il existe aussi de façon plus ponctuelle des surincidences en Bretagne et en Lorraine à certaines périodes.

Dès le premier temps (le mois de janvier 2013), un cluster significatif était détecté sur la côte méditerranéenne,
regroupant une grande partie du Languedoc-Roussillon jusqu'aux Bouches du Rhône. Le

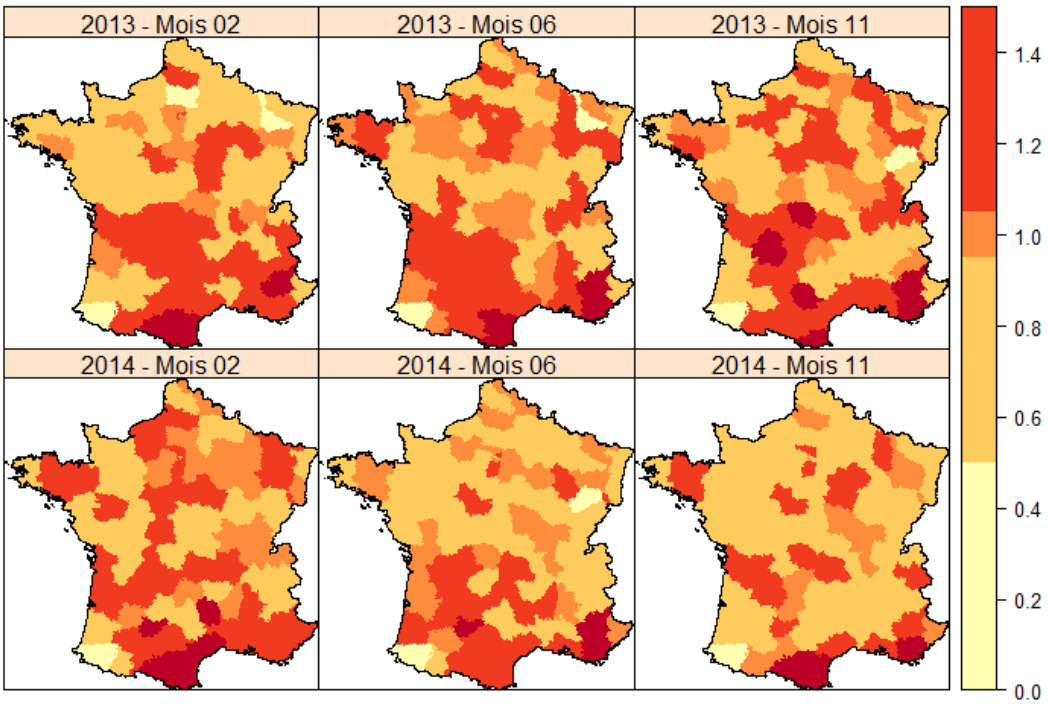

FIG. 19 : Cartes de ratio d'incidence standardisés (SIR) à différents temps

risque relatif par rapport au taux attendu, correspondant à une moyenne nationale, y était de 1,27. A chaque temps suivant, un cluster significatif était observé dans la même zone mais avec une extension territoriale variable. Le risque relatif s'échelonnait de 1,18 lorsque la superficie était la plus étendue, à 1,36 lorsqu'elle était concentrée sur les cinq départements de la moitié sud du Languedoc-Roussillon faisant systématiquement partie du cluster. Le score associé au cluster le plus probable était en augmentation constante à chaque temps de l'analyse, témoignant du fait que le secteur concerné avait un nombre de cas observé systématiquement supérieur à celui attendu après standardisation au cours de la période étudiée. Le tableau de résultats détaillé est consultable en annexe.

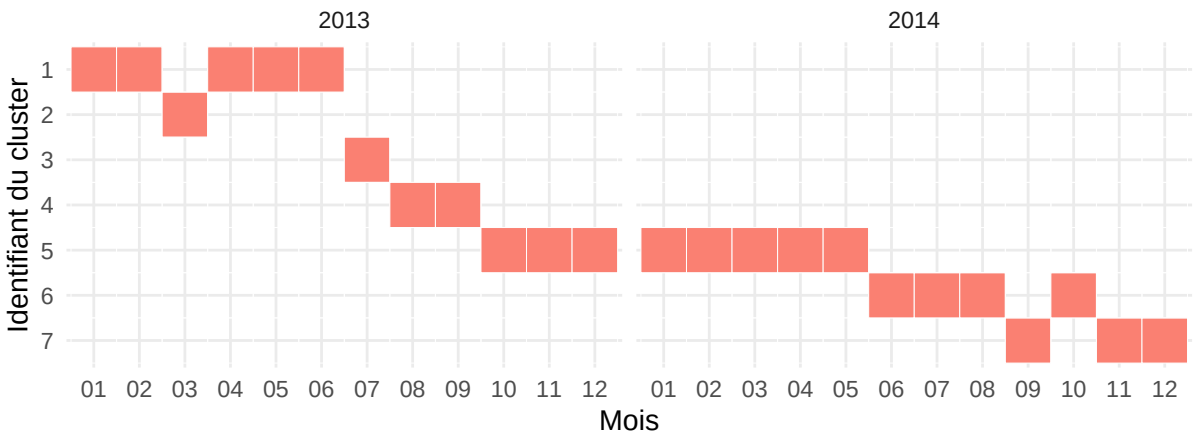

FIG. 20 : Cluster significatif le plus probable à chaque temps de l'analyse, se reporter à la figure suivante pour visualiser la localisation du cluster correspondant

5.4.3 Discussion

Un cluster significatif est identifié à chaque temps de l'analyse, qui varie en taille mais est toujours centré sur le Languedoc-Roussillon, et parfois s'étend jusqu'en Provence-Alpes-Côte d'Azur (PACA). Ce résultat est concordant avec celui d'une étude de la Direction de la Recherche, des Etudes, de l'Evaluation et des Statistiques (DREES) portant sur les IVG réalisées en 2015 (65). Elle identifiait les régions PACA, Île-de-France et Occitanie comme celles ayant le taux de recours le plus élevé en France métropolitaine (> 16 pour mille femmes de 15 à 49 ans, tandis que les autres régions présentaient un taux de 13,5 pour mille au maximum). Ces résultats indiquent une certaine stabilité des comportements en matière de contraception et de recours à l'IVG à l'échelle départementale, mais aussi l'incapacité du système de santé à réduire les inégalités territoriales observées.

Dans une perspective de ciblage territorial, il serait intéressant de considérer une échelle spatiale plus fine, comme l'unité géographique PMSI. Toutefois, l'importance du sur-risque observé à l'échelle de plusieurs départements contigus laisse penser que même avec une meilleure précision spatiale les clusters détectés seraient probablement de grande taille. Utiliser une échelle plus fine pose aussi rapidement un problème de temps de calcul. En

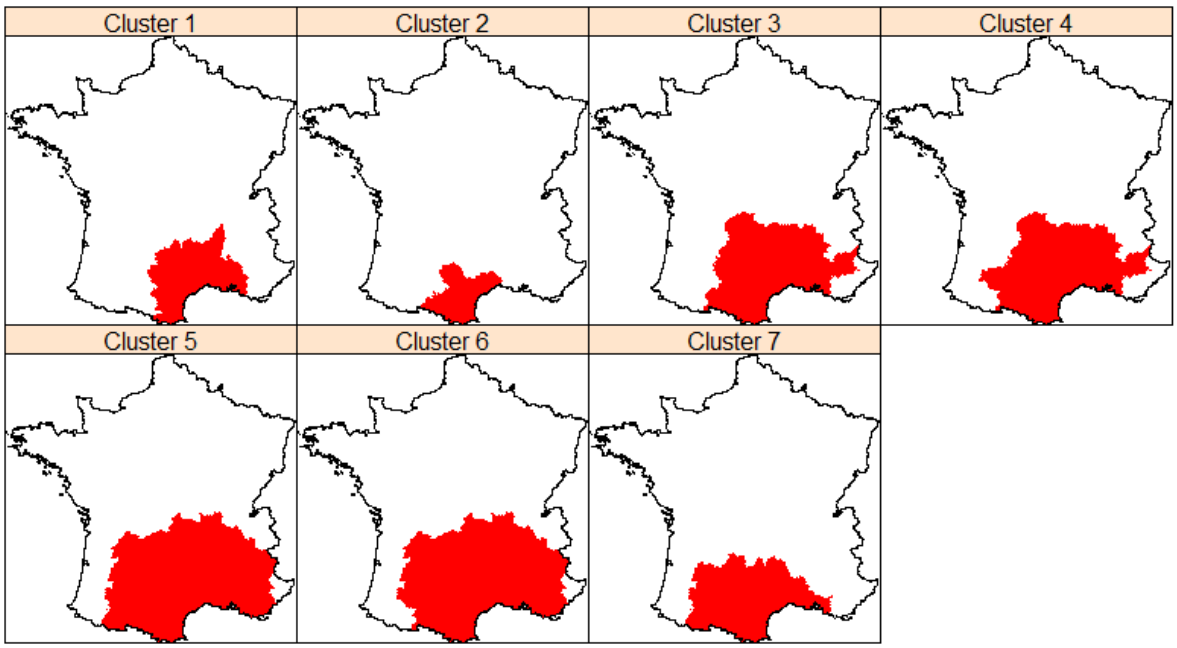

FIG. 21 : Cartes des différents clusters détectés

effet, ce dernier dépend du nombre d'unités de temps et du carré du nombre d'unités spatiales étudiées, ainsi que du nombre de simulations de Monte Carlo réalisées lors de la phase d'inférence. Ainsi, passer d'une centaine de départements à plus de 5500 unités PMSI implique une multiplication par 55^2 soit 3025 du temps de calcul. Le même phénomène est observé si l'on cherche à utiliser un seuil de risque α plus faible afin de mieux contrôler les faux positifs, que ce soit pour la prise en compte de la répétition des analyses dans le temps ou d'une incertitude sur les estimations des attendus. Il faut alors réaliser davantage de simulations de Monte Carlo pour obtenir la précision suffisante, soit un temps de calcul multiplié par 10 pour chaque décimale supplémentaire. Le coût computationnel peut alors rapidement devenir prohibitif, voire dépasser la durée de l'intervalle de temps entre deux analyses successives lorsque l'on dispose de données hebdomadaires voire quotidiennes.

Les facteurs socio-économiques, facteurs favorisants de grossesse non désirée, ne pouvaient être pris en compte dans la mesure où les informations correspondantes n'étaient pas disponibles dans le PMSI. Toutefois, même si cela avait été possible, il aurait été préférable de ne pas en tenir compte dans la mesure où l'objectif n'était pas d'identifier des territoires de surincidence inexpliquée, mais bien de réduire le risque de grossesse non désirée, quel qu'en soient les facteurs favorisants.

L'objectif ici n'était pas de prévenir le recours à l'IVG, mais bien de prévenir les grossesses non désirées. L'IVG en soi n'est qu'un outil de prise en charge de ces grossesses, qui plus est un outil que la population s'est bien approprié comme en atteste le nombre moyen de naissances issues d'une grossesse non désirée par femme de 0,05 alors que le nombre moyen d'IVG est de 0,5 (64). Le fort recours à cette technique dans ces situations en fait un bon marqueur du taux de ces grossesses. S'il existe des disparités territoriales dans son accès, celles-ci ne semblent pas causer de difficultés significatives dans la plupart des cas (67). Toutefois, l'introduction d'une démarche de de prévention n'a d'intérêt que si une partie de ces grossesses est effectivement évitable.

En 2007, seules 37,9 % des femmes ayant recours pour la première fois à une IVG n'avaient pas de moyen de contraception (68), ce qui signifie que la majeure partie des grossesses non désirées sont dues soit au fait qu'aucune méthode de contraception ne soit efficace à 100%, soit à une inadaptation de la méthode au mode de vie des patientes (68). Une étude néo-zélandaise de 2012 a ainsi montré que la prescription de méthodes de contraception au long cours après un premier avortement réduisait le recours à un second IVG (69). Alors qu'en France, depuis les années 90, le nombre d'IVG antérieures n'a presque plus d'impact sur la probabilité d'en réaliser à nouveau (68), une meilleure information des professionnels de santé comme des utilisatrices concernant la diversité des méthodes existantes et l'importance de leur adaptation au mode de vie pourrait permettre une diminution du nombre de grossesses non désirées. Par ailleurs, des études suggèrent une augmentation du nombre d'IVG réalisées chez les mineures dans les années 2000, potentiellement en lien avec le faible recours à la contraception dans cette population, et recommendent un développement des actions d'information (67). Il s'agit d'une part de mieux informer les jeunes femmes sur les aspects pratiques de la contraception (durée d'efficacité, couverture santé, etc.), et d'autre part de meux informer les professionnels de

santé sur les méthodes au long cours (67). Enfin, l'augmentation du nombre de recours multiples à l'IVG observé ces dernières années serait quant à lui davantage dû à l'augmentation de l'âge d'entrée dans un couple stable qu'à des problèmes de contraception, ce facteur étant déterminant dans la désirabilité de la grossesse (68).

En conclusion, les méthodes de détection de cluster peuvent être d'un apport utile pour identifier des cibles d'action privilégiées en matière de prévention et promotion de la santé, mais aussi pour évaluer l'efficacité de ces actions. En effet, si les faibles variations observées sur la question particulière des grossesses non désirées ne justifient pas forcément un suivi mensuel, un suivi régulier est toutefois nécessaire afin d'identifier les territoires où la situation se dégraderait, et également pour vérifier la régression voire la disparition du cluster qui avait conduit au ciblage.

5.5 Conclusion

Les méthodes de statistique de scan conditionnelle et non conditionnelle pour distribution de Poisson ont été appliquées dans plusieurs situations épidémiologiques correspondant à des contextes d'utilisation distincts. Le premier portait sur la détection des épidémies de grippe à partir des données de consultation aux urgences, le deuxième sur la détection de clusters de diffusion active du SARS-Cov-2 dans le cadre du suivi de la pandémie de Covid-19, le troisième sur la détection d'épisodes de surmortalité, et le dernier sur l'identification de territoires présentant un besoin en matière de prévention des grossesses non désirées.

Les méthodes de détection de cluster ont globalement montré leur capacité à identifier les événements sanitaires exceptionnels ou à localiser des territoires prioritaires pour les actions de prévention et promotion de la santé. De plus les territoires visés étaient de toutes les tailles, du département isolé à l'ensemble de la France métropolitaine, ce qui constitue l'intérêt principal de ces méthodes par rapport au simple suivi temporel : elles permettent de détecter des clusters d'événements sans a priori sur leur taille ou leur localisation. De plus, il a été démontré qu'elles étaient supérieures en termes de puissance par rapport au suivi en parallèle de plusieurs unités spatiales (16).

Cependant, malgré l'intérêt potentiel de ces méthodes dans les différents contextes étudiés, quelques limites ont aussi été mises en évidence. Le début des événements sanitaires était détecté rapidement mais, selon le jeu de données étudié, la fin de l'événement survenait soit trop précocément soit trop tardivement. Cela pouvait s'expliquer par un manque de puissance ou au contraire par un taux de faux positifs trop important, autre limite soulevée dans l'exemple du suivi de la grippe. La nature cyclique des phénomènes étudiés contribuait probablement à la difficulté de paramétrer le modèle de façon à optimiser ces deux critères (puissance et taux de faux positifs). Il reste à déterminer la méthode de fixation des effectifs attendus qui permettra de les optimiser dans le contexte de phénomènes périodiques.

Une autre limite apparait lorsque l'on souhaite atteindre une plus grande précision spatiale, ou que l'on souhaite imposer un risque α de faux positifs plus faible. Le problème est

alors le temps de calcul. En effet, ce dernier dépendant à la fois du nombre d'unités de temps, du carré du nombre d'unités spatiales et du nombre de simulations de Monte Carlo, l'augmentation de la précision spatiale, temporelle ou dans l'estimation de la valeur-p entraine un accroissement très important du temps de calcul qui peut rendre difficile voire impossible en pratique la réalisation des analyses de façon répétée.

6 Conclusion générale

Ce travail de thèse portait sur l'application des méthodes statistiques de détection d'agrégats d'événements de santé à la surveillance épidémiologique. La première partie s'appliquait à définir la surveillance épidémiologique et les grands concepts associés. Dans la deuxième partie, les différentes méthodes de détection de cluster proposées dans le cadre prospectif étaient présentées avec leurs particularités et leur contexte d'utilisation. Plusieurs méthodes pouvant être utilisées dans le cadre commun de données agrégées à l'unité spatiale, sous l'hypothèse d'une distribution de Poisson, la troisième partie s'attachait à comparer leurs capacités de détection sur données simulées. La quatrième et dernière partie consistait en l'application de ces méthodes à divers contextes épidémiologiques réels afin d'évaluer l'intérêt éventuel de leur mise en oeuvre dans le cadre d'un système de surveillance.

Au sein de la troisième partie, cinq méthodes de détection de cluster prospectives ont été comparées sur la base de critères de puissance, de sensibilité et de valeur prédictive positive. Les données utilisées avaient été simulées par Kulldorff pour pouvoir procéder à ce type d'études comparatives. Les résultats obtenus permettaient de privilégier la méthode non conditionnelle de Neill lorsqu'il est possible d'estimer le nombre de cas attendu à aprtir de données historiques, à défaut la méthode conditionnelle de Kulldorff lorsqu'il est possible d'estimer la population à risque, et la méthode de scan par permutations sinon. Les deux autres méthodes évaluées présentaient de fortes variations de performances en fonction du risque relatif et du nombre de cas attendu dans le cluster sous l'hypothèse alternative, et nécessitent par conséquent des études supplémentaires afin d'identifier les situations dans lesquelles elles pourraient être préférables aux précédentes. En particulier, la méthode non conditionnelle de Tango présente de faibles puissance et sensibilité mais une bonne spécificité dans la plupart des cas, tandis que la méthode C-CUSUM de Sonesson obtient de fortes puissance et sensibilité lorsque le nombre d'attendus ou que la taille du cluster sont importants, mais au prix d'une plus faible VPP, le cluster détecté ayant tendance à être plus grand que le cluster réel.

Dans la quatrième partie, trois contextes d'utilisation différents ont été testés à partir de données réelles. Le premier consistait en la détection d'épidémies de grippe, le deuxième concernait le suivi de l'épidémie de Covid-19 en France, le troisième était centré sur la détection de clusters de surmortalité et le quatrième sur le ciblage d'actions de prévention des grossesses non désirées sur les territoires en ayant le plus besoin. Dans l'ensemble les méthodes de détection de cluster utilisées ont montré leurs capacités à détecter correctement les événements recherchés. Dans le cas du suivi de la grippe, les épidémies des hivers 2016-2017 et 2017-2018 sont détectées dès leur apparition et leurs localisations étaient concordantes avec celles publiées par Santé Publique France, en revanche des faux positifs apparaissaient entre deux épidémies dans des régions fortement touchées. Les principales régions atteintes par le Covid-19 étaient correctement identifiées dès le premier jour du suivi et tout au long de l'étude, aucun faux positif n'était détecté. Lors de l'application à la surveillance des épisodes de surmortalité, l'épidémie de grippe de l'hiver 2018-2019 était identifiée dès son début, de même que les canicules de juin et juillet 2019 et l'épidémie de Covid-19 à partir de mars 2020. Tous ces épisodes étaient correctement

localisés, toutefois les clusters avaient tendance à rester significatifs au-delà de la durée réelle de l'événement, mais l'objectif était la détection précoce et non l'identification de la fin d'un épisode de surmortalité. Enfin, les clusters de recours à l'IVG détectés dans le cadre de la prévention des grossesses non désirées étaient concordants avec les données de la littérature, mais leur persistance dans le temps indiquait le faible intérêt d'un suivi mensuel, un intervalle de temps plus important pourrait être utilisé.

La principale limite identifiée dans le cadre des applications épidémiologiques concernait le contrôle du risque de faux positif lors du suivi de phénomènes cycliques. En effet, dans ce cas la variation régulière du nombre d'événements attendu entraine une sous-pondération des périodes non-épidémiques dans le calcul de la statistique de test, ce qui diminue la capacité de la méthode à faire redescendre le score sous le seuil de significativité après une épidémie. Dès lors, la détermination de la méthode la plus adéquate pour contrôler les risques de faux positifs tout en préservant la puissance dans ce contexte de phénomènes périodiques est un enjeu important. Une autre limite concernait la capacité de passage à l'échelle de ces méthodes. Ces dernières ont une complexité algorithmique (donc un temps de calcul associé) qui dépend du carré du nombre d'unités spatiales, du nombre d'unités de temps et du nombre de simulations de Monte Carlo réalisées lors de l'inférence statistique. Ainsi, toute augmentation de la précision temporelle, spatiale ou dans l'estimation de la significativité entraine un accroissement important du coût computationnel et du délai nécessaire à l'analyse, ce qui n'est pas toujours compatible avec une application à la surveillance en temps réel d'événements sanitaires.

L'utilisation des méthodes de détection de cluster dans le cadre de la surveillance épidémiologique ouvre de multiples perspectives de recherche. Une première pourrait concerner la recherche d'une méthode d'estimation des attendus qui permet de garantir un contrôle correct des faux positifs en limitant la perte de puissance associée, éventuellement en combinant un léger biais positif dans les estimations avec un seuil de significativité plus strict. Une autre perspective importante en santé publique serait de réduire la complexité algorithmique et donc le temps de calcul associé à ces méthodes, afin de pouvoir étudier en routine des échelles spatiales et temporelles plus fines tout en maitrisant le risque de première espèce, pour cela une solution intéressante serait de pouvoir se passer de simulations de Monte Carlo. Les données utilisées ici pour la détection de cluster étaient des données de santé déjà fréquemment suivies en surveillance épidémiologique (consultations aux urgences, mortalité, nombre d'actes médicaux d'IVG). Il serait intéressant d'élargir les sources de données utilisées dans le domaine de la santé : consommation de médicaments ou de soins ambulatoires qui restent peu exploités en France. Enfin, on constate que les méthodes de détection de cluster restent peu utilisées en routine en raison de leur méconnaissance et de la durée des analyses, mais aussi du manque d'outil d'usage simple pour le non spécialiste. Le développement d'un outil de ce type permettrait de simplifier et développer leur utilisation.

7 Annexes

7.1 Application à la grippe : résultats détaillés

TAB. 4 : Clusters les plus probables à chaque temps t
(de janvier 2017 à décembre 2018)

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
1	1	1	323.73900	1.471346	0.0010
2	2	2	483.82413	1.303802	0.0005
3	3	3	310.97022	1.570757	0.0003
4	3	4	297.86618	1.457679	0.0003
5	3	5	205.53132	1.323390	0.0002
6	4	6	181.82495	10.077048	0.0002
7	4	7	163.46562	8.591577	0.0001
8	4	8	152.94608	7.668524	0.0001
9	4	9	141.24851	6.914408	0.0001
10	4	10	132.14591	6.371225	0.0001
11	4	11	125.18312	5.979653	$< 10^{-4}$
12	4	12	119.96176	5.698852	$< 10^{-4}$
13	4	13	116.12848	5.499399	$< 10^{-4}$
14	4	14	113.37317	5.359426	$< 10^{-4}$
15	4	15	114.76748	5.342772	$< 10^{-4}$
16	4	16	113.39549	5.275296	$< 10^{-4}$
17	4	17	112.45520	5.229431	$< 10^{-4}$
18	4	18	111.81440	5.198349	$< 10^{-4}$
19	4	19	111.37700	5.177213	$< 10^{-4}$
20	4	20	111.07558	5.162687	$< 10^{-4}$
21	4	21	110.86413	5.152515	$< 10^{-4}$
22	4	22	110.71188	5.145200	$< 10^{-4}$
23	4	23	110.59847	5.139757	$< 10^{-4}$
24	4	24	110.51049	5.135537	$< 10^{-4}$
25	4	25	110.43898	5.132108	$< 10^{-4}$
26	4	26	110.37780	5.129177	$< 10^{-4}$
27	4	27	110.32256	5.126531	$< 10^{-4}$
28	4	28	110.26984	5.124007	$< 10^{-4}$
29	4	29	110.21669	5.121463	$< 10^{-4}$
30	4	30	110.16019	5.118760	$< 10^{-4}$

TAB. 4 : Clusters les plus probables à chaque temps t (de janvier 2017 à décembre 2018) *(continued)*

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
31	4	31	110.09708	5.115742	$< 10^{-4}$
32	4	32	110.02337	5.112219	$< 10^{-4}$
33	4	33	109.93384	5.107943	$< 10^{-4}$
34	4	34	109.82154	5.102582	$< 10^{-4}$
35	4	35	109.67705	5.095691	$< 10^{-4}$
36	4	36	109.48773	5.086672	$< 10^{-4}$
37	4	37	109.23694	5.074744	$< 10^{-4}$
38	4	38	108.90328	5.058907	$< 10^{-4}$
39	4	39	108.46018	5.037934	$< 10^{-4}$
40	4	40	107.87596	5.010380	$< 10^{-4}$
41	4	41	107.11459	4.974642	$< 10^{-4}$
42	4	42	106.13729	4.929047	$< 10^{-4}$
43	4	43	104.90478	4.871992	$< 10^{-4}$
44	4	44	103.38015	4.802093	$< 10^{-4}$
45	4	45	101.53167	4.718340	$< 10^{-4}$
46	4	46	102.41121	4.689698	$< 10^{-4}$
47	4	47	99.80361	4.575586	$< 10^{-4}$
48	4	48	113.67592	4.810030	$< 10^{-4}$
49	4	49	127.30763	5.008998	$< 10^{-4}$
50	5	2	778.37672	2.011078	$< 10^{-4}$
51	6	2	2912.90195	2.407532	$< 10^{-4}$
52	7	2	7992.56494	2.644054	$< 10^{-4}$
53	8	4	9465.12202	2.217522	$< 10^{-4}$
54	8	5	6902.01057	1.831535	$< 10^{-4}$
55	8	6	4407.08178	1.549541	$< 10^{-4}$
56	9	8	2335.59003	1.440456	$< 10^{-4}$
57	9	9	1064.82526	1.260411	$< 10^{-4}$
58	10	8	497.86002	5.181881	$< 10^{-4}$
59	10	9	485.21783	4.788848	$< 10^{-4}$
60	10	10	472.11493	4.493728	$< 10^{-4}$
61	10	11	455.48404	4.253733	$< 10^{-4}$
62	10	12	436.26707	4.053718	$< 10^{-4}$

TAB. 4 : Clusters les plus probables à chaque temps t (de janvier 2017 à décembre 2018) *(continued)*

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
63	10	13	428.52606	3.938341	$< 10^{-4}$
64	10	14	416.85143	3.830887	$< 10^{-4}$
65	10	15	410.87101	3.764124	$< 10^{-4}$
66	10	16	412.00185	3.737944	$< 10^{-4}$
67	10	17	411.44768	3.712881	$< 10^{-4}$
68	10	18	411.23163	3.694777	$< 10^{-4}$
69	10	19	408.28473	3.669789	$< 10^{-4}$
70	10	20	406.18682	3.650952	$< 10^{-4}$
71	10	21	405.97264	3.641465	$< 10^{-4}$
72	10	22	404.87234	3.629803	$< 10^{-4}$
73	10	23	404.04906	3.620127	$< 10^{-4}$
74	10	24	403.40510	3.611738	$< 10^{-4}$
75	10	25	402.86658	3.604115	$< 10^{-4}$
76	10	26	402.37591	3.596854	$< 10^{-4}$
77	10	27	401.88711	3.589635	$< 10^{-4}$
78	10	28	400.08763	3.577247	$< 10^{-4}$
79	10	29	398.22856	3.564475	$< 10^{-4}$
80	10	30	397.55527	3.556088	$< 10^{-4}$
81	10	31	395.51228	3.542125	$< 10^{-4}$
82	10	32	393.36277	3.527470	$< 10^{-4}$
83	10	33	393.62011	3.521831	$< 10^{-4}$
84	10	34	391.24860	3.505765	$< 10^{-4}$
85	10	35	391.28180	3.498712	$< 10^{-4}$
86	10	36	388.71442	3.481437	$< 10^{-4}$
87	10	37	387.30764	3.468425	$< 10^{-4}$
88	10	38	384.57139	3.450137	$< 10^{-4}$
89	10	39	381.74662	3.431316	$< 10^{-4}$
90	10	40	378.81497	3.411847	$< 10^{-4}$
91	10	41	376.96627	3.396184	$< 10^{-4}$
92	10	42	373.70157	3.374667	$< 10^{-4}$
93	10	43	370.18013	3.351547	$< 10^{-4}$
94	10	44	371.12715	3.344486	$< 10^{-4}$

TAB. 4 : Clusters les plus probables à chaque temps t (de janvier 2017 à décembre 2018) *(continued)*

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
95	10	45	367.94795	3.320579	$< 10^{-4}$
96	10	46	365.30358	3.297082	$< 10^{-4}$
97	10	47	361.76960	3.268127	$< 10^{-4}$
98	10	48	354.72794	3.223367	$< 10^{-4}$
99	10	49	347.40423	3.174222	$< 10^{-4}$
100	10	50	347.37415	3.147883	$< 10^{-4}$
101	10	51	342.56150	3.098815	$< 10^{-4}$
102	10	52	334.83788	3.033729	$< 10^{-4}$
103	4	103	330.41552	5.593027	$< 10^{-4}$
104	4	104	324.58755	5.485686	$< 10^{-4}$

7.2 Application au Covid-19 : résultats détaillés

TAB. 5 : Clusters significatifs à chaque temps t (du 19 mars au 26 avril 2020)

Temps	Rang	Cluster	Durée	Risque relatif	Score	Valeur-p
1	1	1	1	1.763115	73.50619	0.0001
2	1	2	2	1.821185	115.51828	0.0001
3	1	2	3	1.818345	165.67103	0.0001
4	1	2	4	1.852827	235.80840	0.0001
5	1	2	5	1.812392	289.31033	0.0001
6	1	2	6	1.794686	368.46118	0.0001
7	2	3	1	3.289170	27.46260	0.0001
7	1	2	7	1.771023	438.78709	0.0001
8	2	5	4	2.137129	24.87145	0.0001
8	1	4	6	1.895560	441.73897	0.0001
9	1	6	5	1.748331	452.15491	0.0001
10	1	7	6	1.691806	494.04331	0.0001
11	1	7	7	1.558353	459.03038	0.0001
12	1	7	7	1.537610	450.38537	0.0001
13	1	1	7	1.592833	461.43304	0.0001
14	1	1	7	1.596419	483.73859	0.0001
15	1	8	7	1.573802	501.40290	0.0001
16	1	1	7	1.611995	526.66859	0.0001
17	1	9	7	1.545434	505.11518	0.0001
18	1	10	7	1.593901	529.50155	0.0001
19	1	9	7	1.602199	569.46273	0.0001
20	1	9	7	1.626437	581.89448	0.0001
21	1	11	7	1.688626	543.81023	0.0001
22	1	11	7	1.714508	530.20336	0.0001
23	1	12	7	2.451659	513.64888	0.0001
23	2	13	7	1.572936	32.71195	0.0001
24	2	14	7	1.689833	27.60108	0.0001
24	1	12	7	2.521202	517.04776	0.0001
25	2	15	7	2.429581	24.04776	0.0001
25	1	12	7	2.602727	536.14856	0.0001
26	2	15	7	2.235779	16.89539	0.0001
26	1	12	7	2.654159	515.69011	0.0001
27	2	15	7	2.288526	16.42471	0.0001
27	1	12	7	2.677106	477.51262	0.0001
28	1	16	7	3.125288	458.70127	0.0001

TAB. 5 : Clusters significatifs à chaque temps t (du 19 mars au 26 avril 2020) *(continued)*

Temps	Rang	Cluster	Durée	Risque relatif	Score	Valeur-p
29	1	17	7	3.442672	453.61722	0.0001
30	2	14	7	1.683071	16.39982	0.0001
30	1	17	7	3.437964	407.95120	0.0001
31	1	17	7	3.492306	412.12558	0.0001
31	2	18	6	1.582407	17.11739	0.0001
32	1	17	7	3.290033	338.37763	0.0001
32	2	18	7	1.523054	15.75364	0.0001
33	1	17	7	3.110954	289.91389	0.0001
34	1	17	7	3.029737	256.26247	0.0001
35	1	19	7	1.638455	228.72856	0.0001
36	1	20	7	1.820234	193.63291	0.0001
37	1	21	7	1.636556	165.46737	0.0001
37	2	22	5	2.379344	19.85926	0.0001
38	1	23	7	1.650371	154.72201	0.0001
38	2	22	6	2.311759	20.51522	0.0001
39	1	23	7	1.669846	155.64307	0.0001
39	2	22	7	2.205047	19.64284	0.0001

7.3 Application à la mortalité : résultats détaillés

TAB. 6 : Clusters les plus probables à chaque temps t
(de janvier 2019 à février 2020)

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
1	NA	1	3.537211	1.094441	0.6000
2	1	1	10.891518	1.048237	0.0035
3	2	2	14.215715	1.038299	0.0003
4	3	3	22.053401	1.034595	0.0003
5	4	4	47.197342	1.044963	0.0002
6	5	3	83.081130	1.068092	0.0002
7	5	4	95.341828	1.063154	0.0001
8	6	5	108.191447	1.062921	0.0001
9	5	8	104.050854	1.046605	0.0001
10	6	9	82.920047	1.040991	0.0001
11	7	10	55.100628	1.042579	$< 10^{-4}$
12	7	11	37.128506	1.033326	$< 10^{-4}$
13	8	12	21.925245	1.029823	$< 10^{-4}$
14	9	13	17.740945	1.027139	$< 10^{-4}$
15	9	14	11.420778	1.021005	0.0039
16	10	15	9.311280	1.030472	0.0268
17	10	16	8.968716	1.029021	0.0381
18	10	17	7.082502	1.025067	0.1782
19	10	18	6.687559	1.023728	0.2394
20	10	19	5.368663	1.020738	0.5393
21	10	20	4.057759	1.017611	0.8644
22	NA	7	3.748468	1.139658	0.9166
23	NA	8	3.803396	1.131929	0.9125
24	NA	9	3.723598	1.123350	0.9257
25	NA	2	3.795401	1.266405	0.9191
26	11	1	23.365538	1.123504	$< 10^{-4}$
27	12	2	25.439421	1.059663	$< 10^{-4}$
28	13	3	14.670199	1.038293	0.0003
29	14	4	10.399183	1.040385	0.0146
30	15	1	55.289037	1.134269	$< 10^{-4}$
31	16	7	16.363067	1.067362	$< 10^{-4}$
32	17	7	16.360370	1.030980	$< 10^{-4}$
33	18	9	12.783354	1.042618	0.0017
34	16	10	9.905466	1.043659	0.0243

TAB. 6 : Clusters les plus probables à chaque temps t (de janvier 2019 à février 2020) *(continued)*

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
35	16	11	10.221869	1.042245	0.0190
36	NA	2	7.557145	1.544429	0.1704
37	NA	22	5.963832	1.101861	0.4802
38	NA	23	6.412399	1.103296	0.3756
39	NA	24	6.010493	1.097763	0.4752
40	NA	38	5.479291	1.071230	0.6163
41	NA	39	4.803736	1.065791	0.7887
42	NA	40	4.733644	1.064472	0.8069
43	NA	41	5.362439	1.067800	0.6564
44	NA	42	4.900702	1.063972	0.7741
45	NA	44	4.470018	1.059501	0.8640
46	NA	44	4.870165	1.062201	0.7865
47	NA	45	4.974099	1.062100	0.7650
48	NA	47	4.171584	1.055434	0.9144
49	NA	48	4.390139	1.056216	0.8844
50	NA	49	4.292334	1.054938	0.9014
51	NA	26	4.540117	1.074631	0.8620
52	NA	51	3.900660	1.051169	0.9474
53	NA	52	3.679022	1.049131	0.9741
54	NA	1	5.659233	1.252720	0.6239
55	NA	2	4.395633	1.193415	0.9157
56	NA	55	3.353168	1.045397	0.9941
57	NA	56	3.306511	1.044613	0.9958
58	NA	57	3.532458	1.045660	0.9919
59	NA	58	3.376325	1.044194	0.9958
60	NA	59	2.386837	1.036761	1.0000
61	NA	60	2.100970	1.034156	1.0000
62	NA	61	1.922801	1.032372	1.0000
63	19	1	16.521756	1.500070	0.0001
64	19	2	102.955171	1.930147	$< 10^{-4}$

7.4 Application à la prévention des grossesses non désirées : résultats détaillés

TAB. 7 : Clusters les plus probables à chaque temps t (de janvier 2013 à décembre 2014)

Temps	Cluster	Durée	Score	Risque relatif	Valeur-p
1	1	1	34.72287	1.273964	0.0010
2	1	2	50.04925	1.233152	0.0005
3	2	3	65.61975	1.357016	0.0003
4	1	4	97.29223	1.229116	0.0003
5	1	5	136.14782	1.244216	0.0002
6	1	6	175.95693	1.253728	0.0002
7	3	7	239.95832	1.224078	0.0001
8	4	8	289.27203	1.230014	0.0001
9	4	9	314.98021	1.226475	0.0001
10	5	10	369.95118	1.177436	0.0001
11	5	11	403.65044	1.177776	$< 10^{-4}$
12	5	12	429.70124	1.176001	$< 10^{-4}$
13	5	13	526.67610	1.187144	$< 10^{-4}$
14	5	14	594.94806	1.191877	$< 10^{-4}$
15	5	15	615.30758	1.188090	$< 10^{-4}$
16	5	16	690.11000	1.193176	$< 10^{-4}$
17	5	17	694.48581	1.187896	$< 10^{-4}$
18	6	18	712.40571	1.185925	$< 10^{-4}$
19	6	19	755.27674	1.186490	$< 10^{-4}$
20	6	20	743.54522	1.180549	$< 10^{-4}$
21	7	21	795.93816	1.261069	$< 10^{-4}$
22	6	22	840.89746	1.183390	$< 10^{-4}$
23	7	23	846.43071	1.258018	$< 10^{-4}$
24	7	24	889.16741	1.259345	$< 10^{-4}$

Bibliographie

1. Langmuir AD. The Surveillance of Communicable Diseases of National Importance. The New England Journal of Medicine. janv 1963 ;268(4) :182□92.

2. Astagneau P. Définitions et Concepts. Dans : Surveillance Épidémiologique Principes, Méthodes et Applications En Santé Publique. Lavoisier. Paris ; 2011. (Formation Permanente).

3. Amar E. Quand Un Agrégat de Malformations Rencontre Une Agence de Santé Publique : Une Impression de Déjà-Vu ? Environnement, Risques & Santé. nov 2016 ;15(6) :465□8.

4. La Veille et l'alerte Sanitaires En France. Saint-Maurice : Institut de Veille Sanitaire ; 2011 p. 60.

5. Texier G, Gaudart J, Queyriaux B. Techniques d'analyse Spatiale. Dans : Surveillance Épidémiologique Principes, Méthodes et Applications En Santé Publique. Lavoisier. Paris ; 2011. (Formation Permanente).

6. Langmuir DAD. Communicable Disease Surveillance : Evolution of the Concept of Surveillance in the United States. Proceedings of the Royal Society of Medicine. juin 1971 ;64 :681□4.

7. Thacker SB, Birkhead GS. Surveillance. Field epidemiology. 1996 ;2 :26□50.

8. Eilstein D, Salines G, Desenclos JC. Veille sanitaire : outils, fonctions, processus. Revue d'Épidémiologie et de Santé Publique. oct 2012 ;60(5) :401□11.

9. Paquet C, Coulombier D, Kaiser R, Ciotti M. Epidemic Intelligence : A New Framework for Strengthening Disease Surveillance in Europe. Eurosurveillance. déc 2006 ;11(12) :5□6.

10. CDC. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks : Recommendations from the CDC Working Group. Morbidity and Mortality Weekly Report. mai 2004 ;53(RR05) :1□11.

11. Henning KJ. What Is Syndromic Surveillance ? Morbidity and Mortality Weekly Report. 2004 ;53 :7□11.

12. Josseran L, Fouillet A. Surveillance Syndromique. Dans : Surveillance Épidémiologique Principes, Méthodes et Applications En Santé Publique. Lavoisier. Paris ; 2011. (Formation Permanente).

13. Espino JU, Wagner M, Szczepaniak C, Tsui F-C, Su H, Olszewski R, et al. Removing a Barrier to Computer-Based Outbreak and Disease Surveillance The RODS Open Source Project. Morbidity and Mortality Weekly Report. Centers for Disease Control & Prevention (CDC) ; 2004 ;53 :32□9.

14. Paulos JA. Innumeracy : Mathematical Illiteracy and Its Consequences. London : Penguin Books ; 1990.

15. Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical Methods for the Prospective Detection of Infectious Disease Outbreaks : A Review. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2012 ;175(1) :49□82.

16. Neill DB. Expectation-Based Scan Statistics for Monitoring Spatial Time Series Data. International Journal of Forecasting. juill 2009 ;25(3) :498□517.

17. Shewhart WA. Economic Control of Quality of Manufactured Product. Princeton : Van Nostrand Reinhold ; 1931.

18. Page ES. Continuous Inspection Schemes. Biometrika. 1954 ;41(1/2) :100□15.

19. Ewan WD, Kemp KW. Sampling Inspection of Continuous Processes with No Autocorrelation Between Successive Results. Biometrika. 1960 ;47(3/4) :363□80.

20. Rogerson PA. Formulas for the Design of CUSUM Quality Control Charts. Communications in Statistics - Theory and Methods. Taylor & Francis ; mars 2006 ;35(2) :373□83.

21. Siegmund D. Sequential Analysis : Tests and Confidence Intervals. Springer Science & Business Media ; 1985.

22. Sonesson C. A CUSUM Framework for Detection of SpaceTime Disease Clusters Using Scan Statistics. Statistics in Medicine. 2007 ;26(26) :4770□89.

23. Kenett RS, Pollak M. Data-Analytic Aspects of the Shiryayev-Roberts Control Chart : Surveillance of a Non-Homogeneous Poisson Process. Journal of Applied Statistics. Taylor & Francis ; févr 1996 ;23(1) :125□38.

24. Roberts SW. A Comparison of Some Control Chart Procedures. Technometrics. Taylor & Francis ; août 1966 ;8(3) :411□30.

25. Shiryaev AN. On the Detection of Disorder in a Manufacturing Process. II. Theory of Probability & Its Applications. janv 1963 ;8(4) :402□13.

26. Genin M. Statistiques de scan : théorie et application à l'épidémiologie [thèse de doctorat]. Université du Droit et de la Santé - Lille II ; 2013.

27. Rogerson PA. Monitoring Point Patterns for the Development of SpaceTime Clusters. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2001 ;164(1) :87□96.

28. Piroutek A, Assunção R, Paiva T. SpaceTime Prospective Surveillance Based on Knox Local Statistics. Statistics in Medicine. 2014 ;33(16) :2758□73.

29. Marshall JB, Spitzner DJ, Woodall WH. Use of the Local Knox Statistic for the Prospective Monitoring of Disease Occurrences in Space and Time. Statistics in Medicine. 2007 ;26(7) :1579□93.

30. Assunção R, Correa T. Surveillance to Detect Emerging SpaceTime Clusters. Computational Statistics & Data Analysis. juin 2009 ;53(8) :2817□30.

31. Rogerson PA. Surveillance Systems for Monitoring the Development of Spatial Patterns. Statistics in Medicine. 1997 ;16(18) :2081□93.

32. Kulldorff M. Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2001 ;164(1) :61□72.

33. Neill DB, Moore AW, Sabhnani M, Daniel K. Detection of Emerging Space-Time Clusters. Dans : Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, Illinois, USA : Association for Computing Machinery ; 2005. p. 218□27. (KDD '05).

34. Kulldorff M. Spatial Scan Statistics : Models, Calculations, and Applications. Dans : Glaz J, Balakrishnan N, rédacteurs. Scan Statistics and Applications. Boston, MA : Birkhäuser Boston ; 1999. p. 303□22. (Statistics for Industry and Technology).

35. Takahashi K, Kulldorff M, Tango T, Yih K. A Flexibly Shaped Space-Time Scan Statistic for Disease Outbreak Detection and Monitoring. International Journal of Health Geographics. avr 2008 ;7(1) :14.

36. Neill DB, Moore AW. Rapid Detection of Significant Spatial Clusters. Dans : Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA, USA : Association for Computing Machinery ; 2004. p. 256□65. (KDD '04).

37. Kulldorff M, Heffernan R, Hartman J, Assunção R, Mostashari F. A SpaceTime Permutation Scan Statistic for Disease Outbreak Detection. PLOS Medicine. févr 2005 ;2(3) :e59.

38. Tango T, Takahashi K, Kohriyama K. A SpaceTime Scan Statistic for Detecting Emerging Outbreaks. Biometrics. 2011 ;67(1) :106□15.

39. Allévius B, Höhle M. An Unconditional SpaceTime Scan Statistic for ZIP-Distributed Data. Scandinavian Journal of Statistics. 2019 ;46(1) :142□59.

40. Tango T. On the Recent Debate on the Space-Time Scan Statistic for Prospective Surveillance : On the Recent Debate on the Space-Time Scan Statistic for Prospective Surveillance. Statistics in Medicine. mai 2016 ;35(11) :1927□8.

41. Correa TR, Assunção RM, Costa MA. A Critical Look at Prospective Surveillance Using a Scan Statistic. Statistics in Medicine. 2015 ;34(7) :1081□93.

42. Kulldorff M, Kleinman K. Comments on « A Critical Look at Prospective Surveillance Using a Scan Statistic » by 'T. Correa, M. Costa, and R. Assunção. Statistics in medicine. mars 2015 ;34(7) :1094□5.

43. Burkom HS, Murphy SP, Shmueli G. Automated Time Series Forecasting for Biosurveillance. Statistics in Medicine. 2007 ;26(22) :4202□18.

44. Neill DB, Moore AW, Cooper GF. A Bayesian Spatial Scan Statistic. Dans : Weiss Y, Schölkopf B, Platt JC, rédacteurs. Advances in Neural Information Processing Systems 18. MIT Press ; 2006. p. 1003□10.

45. Kulldorff M, Zhang Z, Hartman J, Heffernan R, Huang L, Mostashari F. Benchmark Data and Power Calculations for Evaluating Disease Outbreak Detection Methods. Morbidity

and Mortality Weekly Report. sept 2004 ;53(Suppl) :144□51.

46. Takahashi K, Tango T. An Extended Power of Cluster Detection Tests. Statistics in Medicine. 2006 ;25(5) :841□52.

47. Sonesson C, Bock D. A Review and Discussion of Prospective Statistical Surveillance in Public Health. Journal of the Royal Statistical Society : Series A (Statistics in Society). 2003 ;166(1) :5□21.

48. Fraker SE, Woodall WH, Mousavi S. Performance Metrics for Surveillance Schemes. Quality Engineering. sept 2008 ;20(4) :451□64.

49. Allévius B. Scanstatistics : Space-Time Anomaly Detection Using Scan Statistics. Journal of Open Source Software. mai 2018 ;3(25) :515.

50. Équipes de surveillance de la gripp. Surveillance de La Grippe En France, Saison 2018-2019. Bulletin Epidémiologique Hebdomadaire. oct 2019 ;2019(28) :552□63.

51. Souty C, Amoros P, Falchi A, Capai L, Bonmarin I, Werf S van der, et al. Influenza Epidemics Observed in Primary Care from 1984 to 2017 in France : A Decrease in Epidemic Size over Time. Influenza and Other Respiratory Viruses. 2019 ;13(2) :148□57.

52. Costagliola D, Flahault A, Galinec D, Garnerin P, Menares J, Valleron AJ. A Routine Tool for Detection and Assessment of Epidemics of Influenza-like Syndromes in France. American Journal of Public Health. American Public Health Association ; janv 1991 ;81(1) :97□9.

53. Neill DB. An Empirical Comparison of Spatial Scan Statistics for Outbreak Detection. International Journal of Health Geographics. avr 2009 ;8(1) :20.

54. Guan W-j, Ni Z-y, Hu Y, Liang W-h, Ou C-q, He J-x, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. New England Journal of Medicine. Massachusetts Medical Society ; avr 2020 ;382(18) :1708□20.

55. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, et al. Early Transmission Dynamics in Wuhan, China, of Novel CoronavirusInfected Pneumonia. New England Journal of Medicine. Massachusetts Medical Society ; mars 2020 ;382(13) :1199□207.

56. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. Nature. Nature Publishing Group ; mars 2020 ;579(7798) :270□3.

57. Salje H, Kiem CT, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, et al. Estimating the Burden of SARS-CoV-2 in France. 2020.

58. Introductions and Early Spread of SARS-CoV-2 in France | bioRxiv. https ://www.biorxiv.org/content/1

59. Šidák Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association. Taylor & Francis ; juin 1967 ;62(318) :626□33.

60. Santé Publique France. COVID-19 : Point Épidémiologique Hebdomadaire Du 16 Avril 2020. 2020.

61. Baghdadi Y, Gallay A, Caserio-Schönemann C, Fouillet A. Evaluation of the French Reactive Mortality Surveillance System Supporting Decision Making. European Journal of Public Health. Oxford Academic ; août 2019 ;29(4) :601□7.

62. Gergonne B, Oza A, Cox B, Wuillaume F, Kaufman Z, Green H, et al. A European Algorithm for a Common Monitoring of Mortality across Europe. p. 42.

63. Insee. Note Méthodologique : Communication Sur La Mortalité Dans Le Cadre de La Pandémie de Covid-19. 2020.

64. Stratégie Nationale de Santé Sexuelle 2017-2030. Ministère des Affaires Sociales et de la Santé ; 2017.

65. Vilain A. Les Interruptions Volontaires de Grossesse En 2015. Etudes et Résultats. juin 2016 ;(0968).

66. Feuille de Route Stratégie Nationale de Santé Sexuelle 2018-2020. Ministère des Solidarités et de la Santé ; 2018.

67. Moisy M. Les IVG chez les mineures : une prise en charge satisfaisante mais une prévention insuffisante. Revue francaise des affaires sociales. La Documentation française ; juin 2011 ;(1) :162□98.

68. Bajos N, Moreau C, Prioux F. Increase of repeat abortion in France : From contraceptive issues to postponement of childbearing age. Epidemiology and Public Health / Revue d'Epidémiologie et de Santé Publique. août 2013 ;61(4) :291□8.

69. Rose SB, Lawton BA. Impact of Long-Acting Reversible Contraception on Return for Repeat Abortion. American Journal of Obstetrics and Gynecology. janv 2012 ;206(1) :37.e1□6.

AUTEUR : Nom : ROUSSELET **Prénom :** Louis

Date de soutenance : 9 juin 2020

Titre de la thèse : Méthodes de détection de clusters spatiaux dans le cadre de la surveillance épidémiologique

Thèse - Médecine - Lille 2020

Cadre de classement : Médecine

DES + spécialité : Santé publique et médecine sociale

Mots-clés : Analyse spatiale ; détection de cluster ; surveillance épidémiologique ; réutilisation de données

Résumé :

Contexte : Plusieurs suspicions de clusters de maladies sont apparues dans des crises sanitaires récentes. Des méthodes statistiques spécifiques ont été développées afin de déterminer si ces agrégats peuvent être expliqués par le hasard ou sont susceptibles de témoigner d'un phénomène réel, mais elles restent peu utilisées dans le cadre de la surveillance épidémiologique. L'objectif de ce travail de thèse était d'étudier la faisabilité de la mise en place d'un système de surveillance épidémiologique orienté vers la détection de clusters spatiaux dans plusieurs contextes d'application : la détection d'épidémies, l'évaluation de leur intensité et le ciblage territorial des actions de prévention / promotion de la santé.

Méthodes : Les caractéristiques essentielles d'un système de surveillance épidémiologique permettant de répondre aux besoins identifiés ont été définies. Au regard de ces caractéristiques, plusieurs méthodes prospectives de détection de clusters spatiaux ont été sélectionnées sur la base d'une revue de la littérature. Une intensive étude de simulation a été conduite pour évaluer et comparer les performances de ces méthodes sur des critères de puissance, de sensibilité et de valeur prédictive positive, en considérant des clusters simulés de taille, de risque relatif et de population variables. Les méthodes sélectionnées à l'issue de cette étape ont ensuite été appliquées dans plusieurs contextes épidémiologiques : la détection d'épidémies de grippe et de clusters de maladie à coronavirus de 2019 (Covid-19), l'évaluation de l'intensité d'événements sanitaires par la mortalité toutes causes et le ciblage territorial des actions de prévention des grossesses non désirées.

Résultats : L'étude de simulation a permis de montrer que les méthodes conditionnelle et non conditionnelle pour distribution de Poisson sont plus performantes et plus stables que les autres méthodes évaluées. Leur application dans différents contextes épidémiologiques a permis de détecter des clusters pertinents en termes de localisation spatiale, après comparaison au gold standard (suivis épidémiologiques publiés). Les épidémies et événements sanitaires exceptionnels ont été détectés précocement après leur apparition.

<u>Composition du Jury</u> :

Président : Professeur P. AMOUYEL

Assesseurs : Professeur A. DUHAMEL, Professeur E. CHAZARD

Directeur de thèse : Docteur M. GENIN