Santé publique France

EXPOSITION ET MILIEUX ENVIRONNEMENTAUX

OCTOBRE 2017

MÉTHODES

ÉVALUATION DE DIFFÉRENTES MÉTHODES DE DÉTECTION D'AGRÉGATS DE CAS DE GASTRO-ENTÉRITES AIGUËS MÉDICALISÉES D'ORIGINE HYDRIQUE

Résumé

Évaluation de différentes méthodes de détection d'agrégats de cas de gastro-entérites aiguës médicalisées d'origine hydrique

Santé publique France mène depuis plusieurs années des travaux reposant sur l'utilisation des données de remboursements de l'Assurance maladie pour améliorer la détection des épidémies de gastro-entérites aiguës médicalisées liées à la consommation d'eau du robinet.

Ce travail s'intègre dans le projet de mise en place d'un système de détection automatisée France entière des agrégats de cas de gastro-entérites aiguës médicalisées (GEAm) d'origine hydrique.

L'intérêt de santé publique pour la recherche automatisée d'agrégats de cas de GEAm réside dans sa portée en prévention du risque à travers l'identification des unités de distribution d'eau (UDI) suspectées d'en être à l'origine. L'identification de ces UDI permettra de déterminer les facteurs de risque et les circonstances à l'origine de cet agrégat.

L'objectif de cette étude était d'évaluer et de comparer les performances de deux méthodes de détection des épidémies de GEA à partir des données de l'Assurance maladie pour identifier les UDI concernées. Les méthodes évaluées sont la statistique de balayage spatio-temporel de Kulldorff et la méthode de comparaison géographique des taux d'incidence.

L'étude de simulation s'est inspirée de la méthode de Noufaily. Trois départements ont été utilisés comme support pour l'analyse : le Puy-de-Dôme, l'Isère et la Gironde. Mille simulations ont été réalisées par département. Les données de l'Assurance maladie ont été utilisées comme données de référence pour la simulation de la ligne de base de la gastro-entérite aigüe.

Ensuite, les méthodes de détection ont été appliquées aux données simulées. La sensibilité et la proportion de fausses alertes ont été utilisées pour comparer les performances des deux méthodes de détection.

Ce travail a permis de choisir une méthode de détection : la statistique de balayage spatio-temporel de Kulldorff. Les résultats de cette étude suggèrent une bonne capacité de détection, quelle que soit la période de l'année pour les épidémies impliquant au moins 10 cas de GEAm attribuables à l'eau. La capacité de détection est très faible en dessous de 5 cas de GEAm attribuables à l'eau.

La méthode retenue sera appliquée sur des données réelles dans le cadre d'une étude pilote et d'un groupe de travail « connexion entre la détection d'agrégats de cas de GEAm et les enquêtes de terrain ». L'identification d'une méthode efficace pour la détection des épidémies d'origine hydrique et l'étude pilote pour tester l'application de la surveillance sur le terrain, incluant la connexion avec les enquêtes environnementales par les ARS, permet d'envisager une surveillance rétrospective des épidémies de GEA hydrique sur l'ensemble du territoire à partir des données de l'Assurance maladie.

MOTS CLÉS : DÉTECTION D'AGRÉGATS, GASTRO-ENTÉRITE AIGUË MEDICALISÉE, ÉPIDEMIE D'ORIGINE HYDRIQUE, STATISTIQUE DE BALAYAGE SPATIO-TEMPOREL, ASSURANCE MALADIE

Citation suggérée : Goria S, Mouly D, Rambaud L, Guillet A, Beaudeau P, Galey C. *Évaluation de différentes méthodes de détection d'agrégats de cas de gastro-entérites aiguës médicalisées.* Saint-Maurice : Santé publique France, 2017. 52 p. Disponible à partir de l'URL : www.santepubliquefrance.fr

Abstract

Evaluation of different methods of detection of aggregates of cases of medicalized acute waterborne gastroenteritis

Santé publique France has been working for several years on the use of health insurance reimbursement data to improve the detection of medicalized acute gastroenteritis (mAGE) outbreaks related to the consumption of tap water.

This work is part of the project to set up an automated detection system in France of aggregates of cases of medicalized acute waterborne gastroenteritis.

The public health interest in the automated search for aggregates of mAGE cases lies in its risk prevention scope through the identification of distribution zones (DZ) suspected of being responsible for them. The identification of these DZ will help to determine the risk factors and the circumstances behind this aggregate.

The objective of this study was to evaluate and compare the performance of two detection methods of AGE outbreaks from Health Insurance data to identify the DZs concerned. The methods evaluated are the Kulldorff space-time scan statistic and the method of geographic comparison of incidence rates.

The simulation study was inspired by the Noufaily method. Three districts were used to support the analysis: Puy-de-Dôme, Isère and Gironde. One thousand simulations per district were carried out. Health Insurance data were used as baseline data for the simulation of the baseline of acute gastroenteritis.

Then, the detection methods were applied to the simulated data. The sensitivity and the proportion of false alarms were used to compare the performance of the 2 detection methods.

This work allowed us to choose a detection method: the Kulldorff space-time scan statistic. The results of this study suggest a good detection capacity, whatever the time of year for outbreaks involving at least 10 waterborne mAGE cases. The detection capacity is very low below 5 cases of mAGE attributable to water.

The method chosen will be applied to actual data in a pilot study and a working group "Connection between the detection of aggregates of mAGE cases and field surveys". The identification of an effective method for the detection of waterborne outbreaks and the pilot study to test the implementation of field monitoring, including the linkage with environmental surveys by the Regional Health Authorities, make it possible to consider a retrospective monitoring of waterborne AGE outbreaks throughout the territory using Health Insurance data.

KEY WORDS: CLUSTER DETECTION, MEDICALIZED ACUTE WATERBORNE GASTROENTERITIS, SPACE-TIME SCAN STATISTIC, HEALTH INSURANCE DATA, WATERBORNE OUTBREAKS

ISSN : EN COURS – ISBN-NET : 979-10-289-0391-6 - RÉALISÉ PAR LA DIRECTION DE LA COMMUNICATION, SANTÉ PUBLIQUE FRANCE — DÉPÔT LÉGAL : OCTOBRE 2017

Auteurs

Sarah Goria[1], Damien Mouly[2], Loïc Rambaud[3], Agnès Guillet[1], Pascal Beaudeau[4], Catherine Galey[5]

[1]Direction santé environnement (DSE), unité traitement, analyse des données et méthodologie, Santé publique France, Saint-Maurice, France

[2]Direction des régions (DiRe), Cire Occitanie, Santé publique France, Toulouse, France

[3]Direction santé environnement (DSE), unité surveillance biologique des expositions et des effets, Santé publique France, Saint-Maurice, France

[4]Direction de la prévention et de la promotion de la santé (DPPS), Santé publique France, Saint-Maurice, France

[5]Direction santé environnement (DSE), unité surveillance des risques et des impacts sanitaires liés aux milieux, Santé publique France, Saint-Maurice, France

Sommaire

Abréviations

ARS	Agence régionale de santé
Cire	Cellule d'intervention en région (Santé publique France)
GEAm	Gastro-entérite aiguë médicalisée
GEA	Gastro-entérite aiguë
Insee	Institut national de la statistique et des études économiques
M_incid	Méthode de comparaison géographique des taux d'incidence
NCA	Nombre de cas attendus
NCO	Nombre de cas observés
RIMref	Variation de référence
RR	Risque relatif
Sniiram	Système national d'information interrégimes de l'Assurance maladie
TIMref	Taux d'incidence médian du secteur de référence
UDI	Unité de distribution d'eau
UDI*	Regroupement de communes correspondant approximativement à une UDI et défini par un algorithme d'optimisation en utilisant les informations liées aux UDI contenues dans la base SISE-Eaux
VI	Variation d'incidence

1. INTRODUCTION

Dans un contexte français caractérisé par une absence de système de surveillance spécifique des épidémies de gastro-entérite aiguë (GEA) d'origine hydrique et une sous-déclaration importante, Santé publique France mène depuis plusieurs années des travaux reposant sur l'utilisation des données de remboursements de l'Assurance maladie pour améliorer la détection des épidémies de GEA hydriques. Des travaux précédents ont permis (i) de définir un cas de GEA médicalisée (GEAm) à partir des informations présentes sur une prescription de médicaments remboursés [1, 2]; (ii) d'évaluer les capacités des données de l'Assurance maladie à décrire une épidémie de GEA hydrique [3-5] ; et (iii) d'élaborer et tester sur des données réelles plusieurs méthodes de détection d'agrégats de cas de GEA potentiellement d'origine hydrique [6].

L'intérêt de santé publique de la recherche automatisée d'agrégats de cas de GEAm réside essentiellement dans sa portée en prévention du risque, non pas à travers sa gestion rapprochée via la restriction de la consommation puisque la détection se fait à distance de l'événement, mais à travers l'identification des Unité de distribution d'eau (UDI) suspectées d'être à l'origine de l'agrégat de cas, l'identification rétrospective des facteurs de risque et des circonstances qui ont entrainé l'épidémie. La découverte d'épidémies doit donc aboutir à une enquête environnementale autour de l'UDI suspectée, et dans l'idéal à l'identification des facteurs techniques, naturels (précipitations…) et humains qui ont prévalu dans l'éclosion de l'épidémie, la définition des mesures de prévention (par l'ARS en lien avec l'exploitant) et leur mise en œuvre (par l'exploitant).

L'objectif de cette étude est d'évaluer et de comparer les performances de deux méthodes de détection des épidémies de GEA hydriques à partir des données de l'Assurance maladie en vue de l'application en routine du système de détection d'agrégats de cas de GEAm spécifique à la surveillance des épidémies d'origine hydrique.

Les méthodes évaluées sont :

- la statistique de balayage spatio-temporel de Kulldorff [7] et
- la méthode de comparaison géographique des taux d'incidence [6].

La première partie de ce rapport présente les méthodes de détection retenues, la deuxième le protocole de simulations. La troisième partie présente les résultats.

2. DESCRIPTION DES MÉTHODES DE DÉTECTION

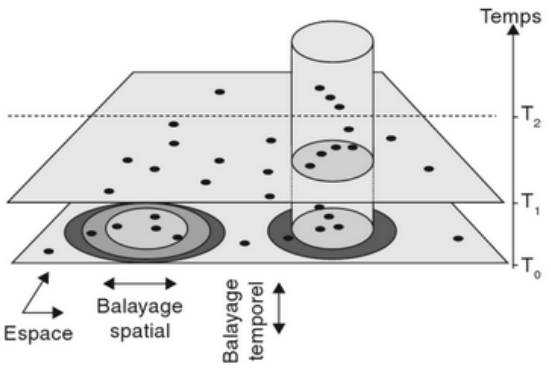
2.1. La statistique de balayage spatio-temporel de Kulldorff

Parmi les méthodes de détection d'agrégats spatio-temporels existantes, la méthode développée par Kulldorff [7] est apparue comme pertinente pour répondre à la problématique de détection des épidémies de gastro-entérite aigüe d'origine hydrique. Il s'agit de la méthode de détection spatio-temporelle de référence à l'heure actuelle. Elle permet de prendre en compte les aspects temporels (prise en compte de l'épidémie hivernale) et les tests multiples. Elle bénéficie de l'existence d'un logiciel gratuit[1] et offre par ailleurs la possibilité d'inclure des cofacteurs dans l'analyse.

La méthode de « balayage spatio-temporel» de Kulldorff [7, 8] permet d'identifier les zones ayant des excès de cas dans l'espace et le temps. Pour une détection uniquement spatiale, cette méthode consiste à réaliser un balayage de toute la zone d'étude par le déplacement d'une fenêtre de forme prédéfinie (disque) de taille variable : cette fenêtre dite glissante, est placée successivement au centroïde de chaque unité géographique (commune) de la zone d'étude et le nombre de cas dans cette fenêtre est comparé au nombre de cas dans l'ensemble de la zone située à l'extérieure de la fenêtre. Cette fenêtre est de taille variable : elle peut s'agrandir en incluant d'autres unités jusqu'à une taille maximale définie *a priori*. Pour une détection spatio-temporelle, une fenêtre alors cylindrique parcourt le temps et l'espace de telle façon que l'ensemble des unités géographiques, des tailles et des durées soient successivement considérées (Figure 1). Finalement, un très grand nombre de fenêtres plus ou moins superposées, de tailles différentes couvrant l'ensemble de la région et de la période d'étude est créé et chacune est candidate pour être un agrégat ou *cluster*.

I FIGURE 1 I

Principe du scan spatial et spatio-temporel de Kulldorff, les points noirs représentant des cas (source: Texier *et al.* [9])



Un agrégat est détecté lorsque le risque à l'intérieur de la fenêtre est significativement supérieur à celui en dehors de cette fenêtre. La statistique de test est fondée sur le rapport de vraisemblances i.e. le rapport de la vraisemblance calculée sous l'hypothèse alternative (le risque à l'intérieur de la fenêtre est supérieur à celui à l'extérieur) et de la vraisemblance calculée sous l'hypothèse nulle d'égalité des risques. La fenêtre ayant le rapport de

[1] http://www.satscan.org

vraisemblances le plus élevé définit l'agrégat le plus probable, i.e. celui qui a le moins de chance de survenir par hasard. Des agrégats secondaires avec un rapport de vraisemblance élevé peuvent aussi être identifiés [8]. Ces agrégats sont d'intérêt parce qu'ils peuvent rejeter à eux seuls l'hypothèse nulle [10].

Les paramètres utilisés

Les données de l'Assurance maladie sont disponibles à la commune de résidence des cas de GEAm et par jour [1]. L'unité spatiale considérée est alors la commune.

Nous avons testé le scan de permutation spatio-temporelle de Kulldorff qui peut être utilisé pour la détection spatio-temporelle d'épidémie en l'absence de références démographiques [7]. Le nombre de cas attendus est alors calculé en utilisant les cas observés. Le nombre de cas attendus $\mu_{z,}$ dans une région Z à l'instant t dépend du nombre de cas observés sur Z dans l'ensemble de la période, ainsi que du nombre de cas observés pour l'ensemble de la région à l'instant t :

$$\mu_{Z,t} = \frac{1}{N}\left(\sum_{Z}\mu_{Z,t}\right)\left(\sum_{t}\mu_{Z,t}\right)$$

où N désigne le nombre total de cas observés sur toute la région durant toute la période. Le logiciel utilisé est SaTScan [10].

La méthode de Kulldorff est appliquée avec deux matrices de voisinage différentes. La première matrice considère comme voisines les communes partageant une même Unité de Distribution d'eau (UDI). Cette matrice est obtenue en exploitant les données des UDI contenues dans la base SISE-Eaux (liste des communes par UDI, la population des UDI, la population des quartiers de commune desservis par une même UDI) et un algorithme qui permet d'optimiser les regroupements de communes [11]. Les critères de décision de l'algorithme sont basés sur la proportion de la population desservie par une unité de distribution d'eau dans chaque commune. Les nouveaux regroupements de communes, UDI*, peuvent correspondre aussi bien à une seule commune (cas d'une commune ne partageant pas d'unité de distribution avec les communes voisines) qu'à un groupe de plusieurs communes (cas de communes partageant une même unité de distribution d'eau). Les signaux identifiés ont en commun le fait de partager la même exposition à l'eau du robinet.

La deuxième matrice est obtenue à partir de la distance : deux communes sont voisines si la distance entre leurs centroïdes est <= 20 km. Le référentiel géographique communal BD TOPO® de l'IGN a été utilisé pour déterminer le centroïde des communes. Cette deuxième matrice est utilisée parce qu'elle permet de s'affranchir des données de la base SISE-Eaux.
L'unité temporelle est le jour. Les variables jours de la semaine et jours fériés sont utilisées lors de la détection.

2.2. La méthode de comparaison géographique des taux d'incidence

La méthode de comparaison géographique des taux d'incidence (M_incid) est construite par la juxtaposition de deux méthodes directement inspirées par la pratique de terrain : la comparaison du taux d'incidence entre la commune ciblée et le reste du département ; la comparaison de l'évolution temporelle du taux d'incidence entre la commune ciblée et le reste du département. La juxtaposition de ces deux informations a pour objectif d'augmenter la spécificité des agrégats détectés [6] : on retient in fine les événements qui conjuguent un taux d'incidence anormalement élevé et une dynamique d'apparition plus rapide que la tendance temporelle départementale.
Chacune des méthodes est mise en œuvre en deux étapes successives pour chaque semaine (du lundi au dimanche) et chaque commune testée : 1) détermination du nombre de cas attendus (NCA), avec des modalités propres à chaque méthode, et 2) test de comparaison du nombre de cas observés (NCO) et du NCA.

Pour la détermination du NCA, l'intégralité du département d'appartenance de la commune testée est utilisée comme secteur de référence :

-Dans la méthode de comparaison spatiale des taux d'incidence, le NCA est le produit du taux d'incidence médian du secteur de référence (TIMref) par la taille de la population de la commune testée. L'utilisation de la médiane permet de limiter l'effet des valeurs extrêmes (e.g. liées aux épidémies hivernales) ou aberrantes. Le TIMref peut cependant être nul, notamment en été et dans les départements ruraux, où la majorité des communes sont de petite taille. Cet inconvénient a été traité par l'exclusion des communes de taille inférieure à 500 habitants dans le calcul du TIMref.

-Dans la méthode de comparaison de l'évolution temporelle de l'incidence, une variation du taux d'incidence entre la période témoin et la période cible est calculée pour chaque commune du secteur de référence. Cette variation correspond au ratio entre 1) le taux d'incidence des cas de GEAm sur la période cible et 2) le taux d'incidence des cas de GEAm sur une période témoin. La variation de référence (RIMref) correspond à la médiane de cet ensemble de valeurs. Le NCA est le produit de RIMref et du nombre de cas observés sur la période témoin sur la commune testée. Le RIMref est calculé en excluant les communes de moins de 500 habitants. La période témoin choisie correspond aux quatre semaines situées entre la cinquième semaine (S-5) et la deuxième semaine (S-2) précédant la semaine cible (S0). Une semaine tampon est ainsi interposée entre la semaine cible et la période témoin afin d'éviter d'inclure dans la période témoin le début d'une éventuelle épidémie. En période de faible incidence (printemps, été), lorsque le nombre de cas observé est nul durant la période témoin dans les communes de petite taille, la valeur 1 est imputée afin de permettre le calcul des variations des taux d'incidence communaux. Cet artifice tend à réduire le nombre d'agrégats détectés.

La détection des agrégats hebdomadaires est fondée sur la comparaison du NCO et du NCA. Les agrégats sont retenus lorsque la significativité (p-value) du test exact reposant sur l'hypothèse d'une distribution poissonienne est strictement inférieure à 1.10^{-5} pour les deux tests réalisés (un pour chaque méthode simple) [6]. Ceux survenus sur une même commune et consécutifs dans le temps sont assemblés en un seul agrégat, d'une durée d'une semaine ou plus. Le nombre de cas et le RR d'un agrégat de cas sont recalculés à partir du cumul des NCO et NCA des agrégats hebdomadaires les constituant.

3. PROTOCOLE DE SIMULATION

L'objectif est d'évaluer la performance des méthodes retenues pour détecter une épidémie de GEAm de type hydrique, à partir de données simulées.

La première partie de ce protocole présente la construction des simulations et la deuxième détaille les critères utilisés pour évaluer et comparer les performances des deux approches de détection.

3.1 Les simulations

La simulation porte sur les données de GEAm. Pour ce protocole de simulations nous nous sommes inspirés de l'article de Noufaily *et al.* [12]. Noufaily *et al.* ont développé un algorithme de détection d'événements inhabituels pour plusieurs maladies infectieuses et en étudient les performances sur la base de simulations.

Nous avons simulé indépendamment les données de référence (qui incluent les épidémies hivernales de GEAm principalement liées à une contamination de personnes à personnes) et les épidémies de GEAm d'origine hydrique. Les épidémies d'origine hydrique ont été ajoutées aux données de référence.

Zone et période d'étude
Les simulations sont réalisées sur les départements du Puy de Dôme, de l'Isère et de la Gironde et s'appuient sur les données des quatre années de 2010 à 2013. Les départements choisis ont des problèmes connus de pollutions microbiologiques chroniques et ont connu chacun une ou plusieurs épidémies d'origine hydrique ayant fait l'objet d'une investigation par Santé publique France. Par ailleurs ils présentent des situations contrastées en matière de schéma de distribution de l'eau.

Indicateur sanitaire et source de données
L'indicateur sanitaire utilisé correspond au nombre de cas de GEAm par jour et par commune identifié à partir des données du Système national d'information inter-régimes de l'Assurance maladie (Sniiram) au moyen d'un algorithme spécifique (consultation d'un médecin et remboursement de médicaments spécifiques prescrits [1, 3]).

Le Sniiram regroupe l'ensemble des prescriptions de soins soumis à remboursement, incluant les médicaments. Le Sniiram couvre près de 99% de la population résidant en France [13].

L'algorithme de discrimination des cas de GEAm a été évalué [1] sur la base d'une collecte de cas symptomatiques avérés (définition : Majovicz [14]) médicalisés. La sensibilité de la définition de cas médicamenteuse est ainsi estimée > 0,80 et la valeur prédictive positive > 0,80.

Simulation des données de référence
Noufaily *et al.* [12] simulent les données de référence en utilisant une loi binomiale négative de moyenne μ(t) et de variance ϕ *μ(t) où ϕ est le paramètre de dispersion. Noufaily *et al.* considèrent le modèle binomial négatif de moyenne :

$$\mu(\mathbf{t}) = \mathbf{exp}\left\{\boldsymbol{\theta} + \boldsymbol{\beta}\mathbf{t} + \sum_{\mathbf{j=1}}^{\mathbf{m}}\left[\boldsymbol{\gamma}_{\mathbf{1\_j}}\,\mathbf{cos}\left(\frac{\mathbf{2\pi jt}}{\mathbf{52}}\right) + \boldsymbol{\gamma}_{\mathbf{2\_j}}\,\mathbf{sin}\left(\frac{\mathbf{2\pi jt}}{\mathbf{52}}\right)\right]\right\}$$

et créent les simulations à partir de différentes combinaisons des paramètres $\theta, \beta, \gamma_{1\_j}, \gamma_{2\_j}$ et ϕ pour avoir différentes tendances, saisonnalités, dispersions et couvrir la variété des données de surveillance de la *Health Protection Agency* britannique.

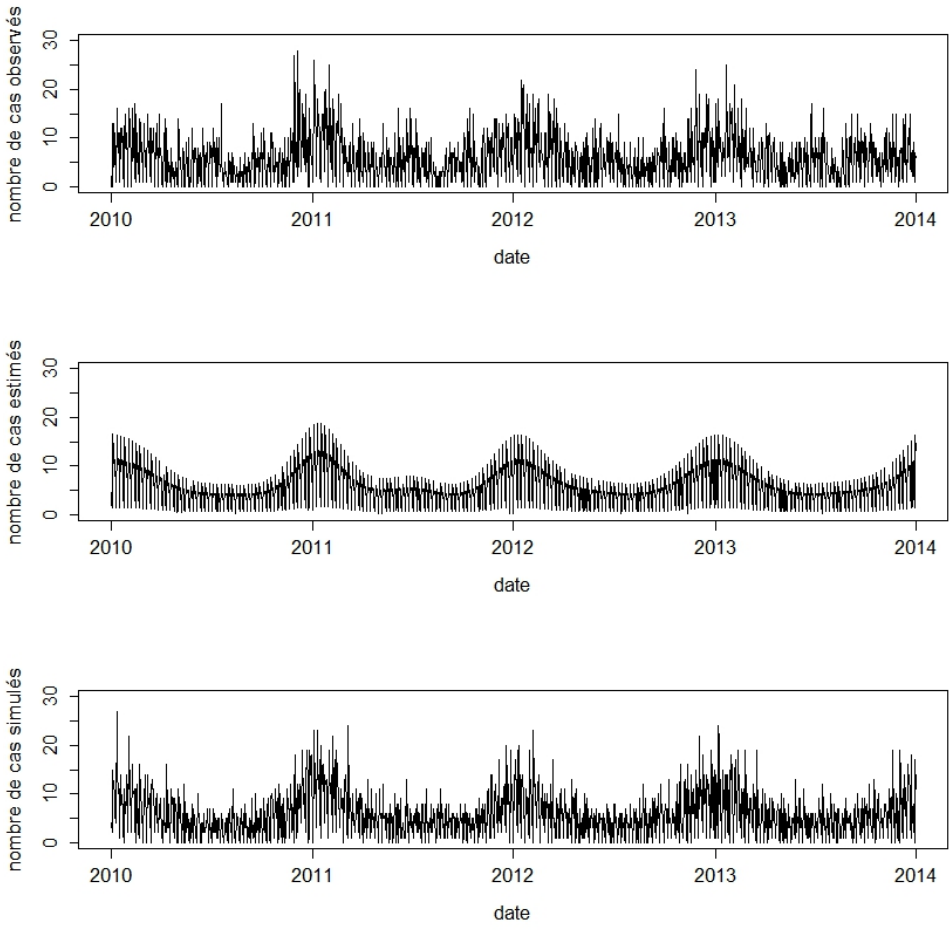
Pour simuler les données de référence de GEAm, nous avons utilisé les cas identifiés à partir du Sniiram pour les trois départements sélectionnés après « lissage » du nombre de cas observé quotidiennement. Pour cela nous avons estimé un nombre de cas attendus par un modèle de Poisson de paramètre $\mu(t)$. Le modèle est estimé en utilisant les données journalières des années 2010 à 2013 agrégées au département. Les variations saisonnières de l'incidence changeant d'une année à l'autre (intensité et position du pic hivernal), un ajustement plus souple que celui proposé par Noufaily nous a paru préférable. La tendance et les variations saisonnières sont ainsi modélisées à l'aide d'une fonction spline pénalisée du temps. Nous avons également pris en compte les variables jours de la semaine et jours fériés. Le nombre de cas attendus estimé au niveau du département est ensuite réparti sur les différentes communes selon leur taille en termes de nombre de cas observés.

Une loi binomiale-négative de moyenne égale au nombre de cas attendus par commune et par jour est ensuite utilisée pour simuler le nombre de cas par jour.

La Figure 2 montre un exemple de simulation des données de références de GEAm.

I FIGURE 2 I

Nombre de cas observés (graphe du haut), nombre de cas attendus estimé (graphe du milieu) et nombre de cas simulé (graphe du bas) pour la commune de code Insee 38553 (population = 18 541 habitants et nombre de cas observés sur la période étudiée = 8 677)

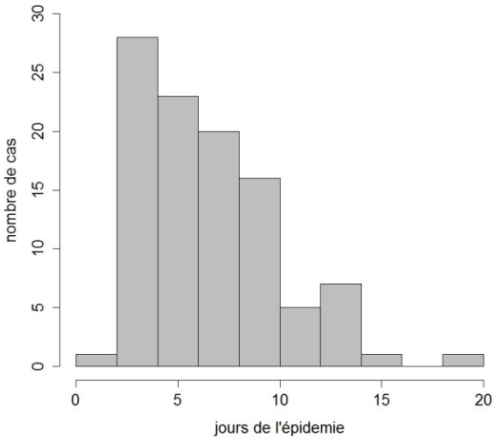


Simulation des épidémies de GEAm de type hydrique

Nous avons simulé des épidémies de GEAm à partir de taux d'incidence observés dans des épidémies de GEA hydrique ayant fait l'objet d'investigations : la taille de l'épidémie est définie à partir de la variation d'incidence de GEAm souhaitée et le nombre de cas est distribué en utilisant une loi log-normale suivant Noufaily *et al.* [12]. La variation d'incidence est le rapport de la différence entre le nombre de cas de l'épidémie et le nombre de cas attendus (référence) et la population. Elle est utilisée pour calculer la taille de l'épidémie simulée (nombre de cas épidémiques). On a défini une loi log-normale dont la médiane se trouve entre 1/3 et 1/2 de la durée de l'épidémie et l'écart type de la loi normale étant fixé à 0.5. Les Figures 3 et 4 présentent des exemples d'épidémies simulées.

I FIGURE 3 I

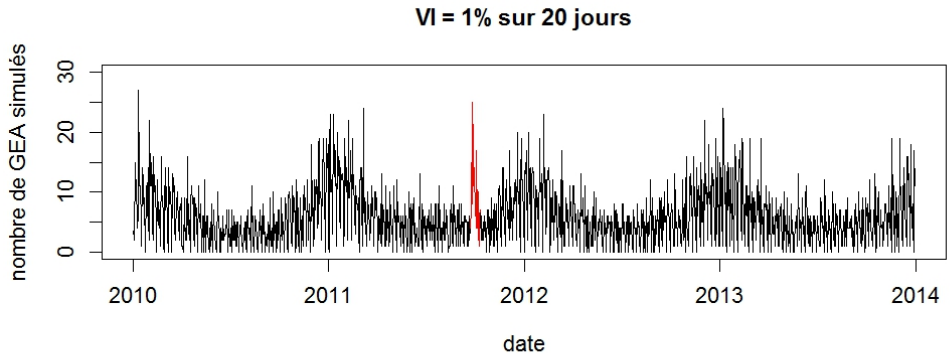
Exemple d'une épidémie simulée sur une commune de l'Isère de 6 377 habitants

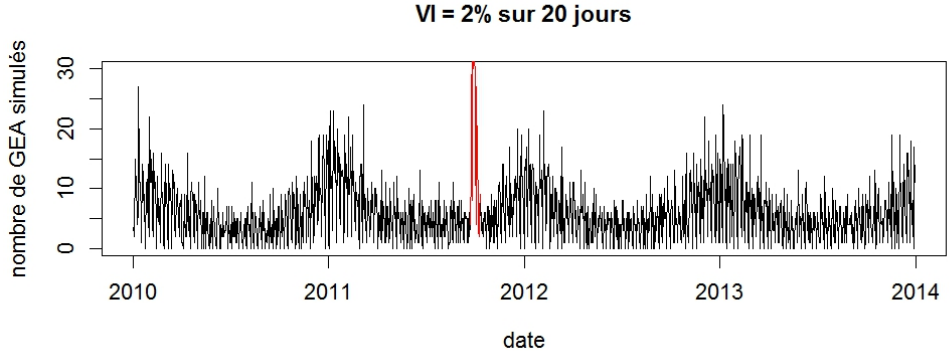


L'histogramme présente le nombre de cas simulés par jour (avec une durée de l'épidémie de 20 jours et une médiane à 6 jours). La taille de l'épidémie est de 127 cas avec une variation d'incidence de 2%

I FIGURE 4 I

Deux épidémies simulées le 22/09/2011 avec une variation d'incidence (VI) de 1% et de 2% d'une durée de 20 jours sur la commune de code Insee 38553 (18 541 habitants)





Les paramètres des simulations des épidémies de GEAm de type hydrique

L'unité géographique d'intérêt pour la simulation des épidémies est l'UDI*. En effet, en faisant l'hypothèse que l'ensemble de l'UDI* est atteinte par la pollution, la population alimentée par une même UDI* peut être supposée semblable du point de vue de l'exposition. Pour cela, pour les simulations, nous partirons de l'UDI*.

L'UDI* concernée par l'épidémie simulée est choisie aléatoirement. Si l'UDI* choisie dessert plusieurs communes, l'excès de cas est réparti sur l'ensemble des communes partageant l'UDI* au prorata des effectifs desservis par cette UDI*. Pour cela nous avons utilisé la base SISE-Eaux, et notamment la liste des communes par UDI et la population des UDI.

Nous n'avons pas fait de simulations sur les UDI* de moins de 200 habitants.

La date de début de l'épidémie est choisie aléatoirement, ainsi que la durée de l'épidémie et la variation d'incidence.

La durée de l'épidémie

Une épidémie de GEA par contamination hydrique peut varier de quelques jours à 4 semaines [15]. La durée de l'épidémie est choisie aléatoirement entre 3 et 28 jours.

La variation de l'incidence de l'épidémie

Une épidémie de GEA hydrique peut être plus ou moins intense et le nombre de cas de GEAm peut varier en fonction de nombreux paramètres (impact de l'épidémie, accès aux soins, habitudes de consultations) [5]. La variation d'incidence est choisie aléatoirement entre 0.5 et 6%.

Afin de tenir compte de l'effet jour dans les données du Sniiram, lié aux périodes de fermeture des structures de soins (médecins, pharmacies), les cas simulés attribués à un samedi ou un dimanche sont reportés sur les premiers jours de la semaine suivante (lundi, mardi). Une règle de report basée sur la répartition des cas de GEAm observés en fonction des jours de la semaine est appliquée (50% des cas simulés le dimanche sont reportés sur les lundi et mardi suivants et 25% des cas simulés le samedi sont reportés sur le lundi).

Nombre de simulations

On a réalisé 1 000 simulations par département sur l'ensemble de la période (3 000 simulations au total).

3.2 Indicateurs pour l'évaluation et la comparaison des méthodes de détection

Pour comparer les performances des deux méthodes de détection, nous avons utilisé :

-la sensibilité définie comme le nombre d'épidémies simulées détectées sur le nombre total d'épidémies simulées et qui mesure la capacité de chacune des méthodes à détecter les épidémies de GEAm simulées et ;

-la proportion de fausses alertes définie comme le nombre d'épidémies détectées non simulées sur le nombre total d'épidémies détectées et qui mesure la capacité de chacune des méthodes à ne pas détecter d'agrégats non simulés (= 1-valeur prédictive positive).

Pour évaluer la sensibilité des méthodes retenues à détecter un agrégat spatio-temporel simulé nous pouvons choisir entre deux approches :

-la première considère qu'un agrégat est détecté si la méthode détecte au moins un jour (méthode de Kulldorff) ou une semaine (méthode M_incid) et une commune appartenant à l'épidémie simulée et l'ensemble correspond à un vrai positif. Cette approche est plutôt « optimiste » notamment dans le cas d'événements durables et qui concernent un regroupement de communes (une UDI*) puisqu'il suffit de détecter un seul jour (ou semaine) et une seule commune pour considérer l'ensemble de l'événement comme détecté.

-la deuxième approche considère que la méthode doit détecter toute la période et toutes les communes appartenant à l'agrégat simulé. Chaque jour et commune détecté appartenant à l'agrégat simulé correspond à un vrai positif [16]. Cette approche est « exigeante » car pour obtenir une sensibilité élevée il faut détecter l'ensemble des jours et communes de l'événement simulé.

Nous avons choisi d'utiliser la première compte tenu que les agrégats identifiés sont ensuite expertisés par les acteurs locaux. Le contour spatio-temporel de l'épidémie – le cas échéant – est ré-estimé sur la base de critères locaux.

La comparaison des méthodes sur la proportion de fausses alertes est plus délicate. La méthode de Kulldorff identifie des agrégats de cas qui peuvent comporter une ou plusieurs communes, comprises dans les limites des UDI* ou dans un rayon de 20km. M_incid identifie des agrégats de cas sur chaque commune prise séparément. Les deux approches suivantes ont été retenues pour calculer la proportion de fausses alertes :

- La première approche définit la proportion de fausses alertes comme le nombre d'agrégats détectés ne correspondant pas à une épidémie simulée (c'est-à-dire ne partageant pas au minimum une journée et une commune avec une épidémie simulée) sur le nombre total d'agrégats détectés.

- La seconde considère chaque commune de l'agrégat détecté séparément comme appartenant ou pas à une épidémie simulée. Toute commune détectée mais n'appartenant pas à une épidémie simulée est une fausse alerte. Le dénominateur est le nombre total de commune/période correspondant à tous les agrégats détectés.

La première approche (critère « agrégats ») permet de comparer les deux variantes de la méthode de Kulldorff pour lesquelles l'emprise des agrégats découverts peut couvrir plusieurs communes. M_incid ne détectant que des agrégats communaux, la deuxième approche a été considérée (critère « communes ») pour comparer M_incid et Kulldorff.

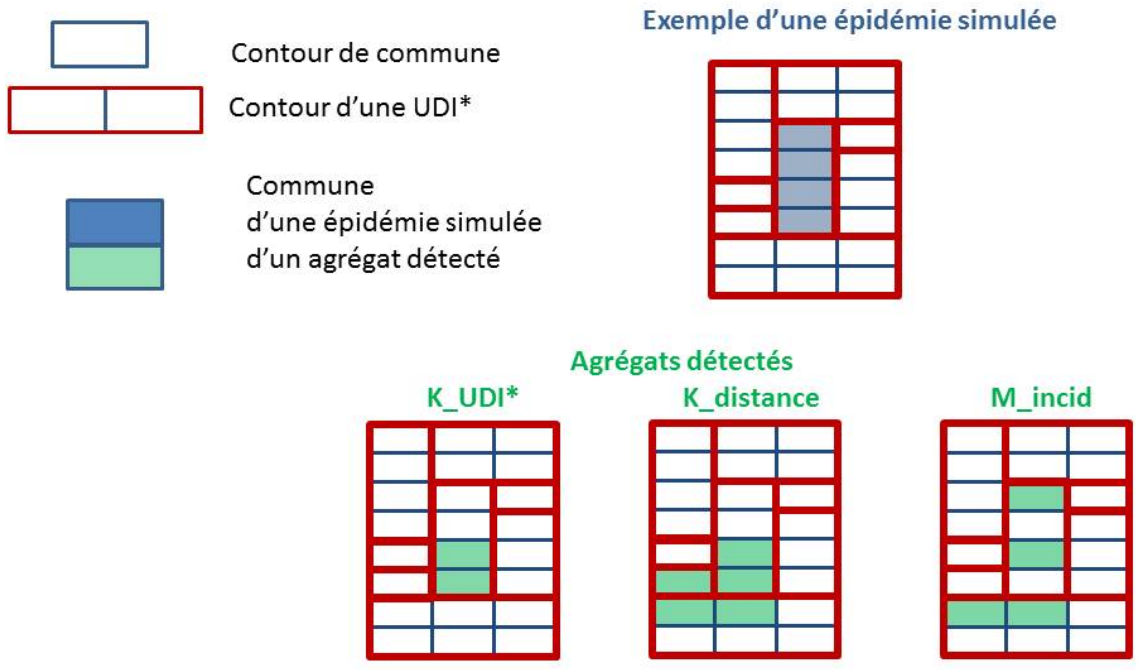
Le critère « commune » permet une comparaison de la proportion de fausses alertes moins pénalisante pour M_incid. Ce critère est intéressant aussi pour comparer les 2 variantes de la méthode de Kulldorff. En effet, la méthode basée sur les UDI contraint les agrégats de s'inscrire dans le périmètre d'une UDI contrairement à la méthode basée sur la distance. Cette dernière peut ainsi correctement identifier une épidémie simulée (il en trouve une partie) mais engendrer autant de fausse alertes que de communes de l'agrégat détecté qui débordent du périmètre de l'UDI* épidémique (Figure 5).

Pour l'analyse et l'interprétation des résultats nous avons regardé les performances des méthodes en fonction de la taille des UDI* simulés (population et nombre de cas simulé) et de la saison (hiver/ hors hiver). La période hivernale a été définie en fonction de l'épidémie de GEA hivernale à transmission inter-humaine qui s'étendait entre début décembre et fin mars pour les années 2010 à 2013. La période hors hiver correspond aux autres mois de l'année (avril à novembre).

Pour la méthode de Kulldorff des indicateurs supplémentaires seront calculés : le nombre de communes détectées appartenant à l'épidémie simulée sur le nombre de communes de l'épidémie simulée et le délai de détection, défini comme le délai entre le début de l'épidémie simulée et le premier jour de l'agrégat détecté par la méthode. Ces indicateurs ne peuvent être calculés pour la méthode M_incid, celle-ci étant définie à la semaine et à la commune.

I FIGURE 5 I

Fausses alertes identifiées par la méthode de Kulldorff (avec UDI*, K_UDI*, et distance, K_dist) et M_incid selon les critères « agrégats » et « communes »



	K_UDI*	K_distance	M_incid
Nbre d'agrégats détectés:	1	1	4
Fausses alertes , exprimées en nbre :			
d'agrégats	0	0	2
de communes	0	3	2

4. RÉSULTATS

4.1 Description des départements étudiés

La description détaillée des départements étudiés en termes de nombre d'habitants, nombre de communes et d'UDI est présentée en Annexe 1.

Pour l'Isère nous avons retenu les 457 UDI de plus de 200 habitants représentant 98% de la population du département. La population moyenne est de 2562 habitants desservis. Les 85 UDI partagées desservent de 2 à 13 communes. Pour le Puy de Dôme nous avons retenu les 172 UDI de plus de 200 habitants représentant 97% de la population du département. La population moyenne est de 3 534 habitants desservis. Les 66 UDI partagées desservent de 2 à 67 communes. Pour la Gironde nous avons retenu les 128 UDI de plus de 200 habitants représentant 99.98% de la population du département. La population moyenne est de 11 100 habitants desservis. Les 68 UDI partagées desservent de 2 à 34 communes.

4.2 Description des simulations

La distribution de la population des UDI* utilisées dans les simulations est présentée dans le Tableau 1. La distribution du nombre de cas simulés est présentée dans le Tableau 2. La distribution de la durée des épidémies simulées est présentée dans la Figure 6.

I TABLEAU 1 I

Distribution de la population des UDI* tirées au sort pour les simulations d'épidémies

	min	P25	médiane	moyenne	P75	max
Isère	1	11	23	70	60	7 392
Puy de Dôme	2	10	22	122	69	5 551
Gironde	2	45	119	341	288	10 150

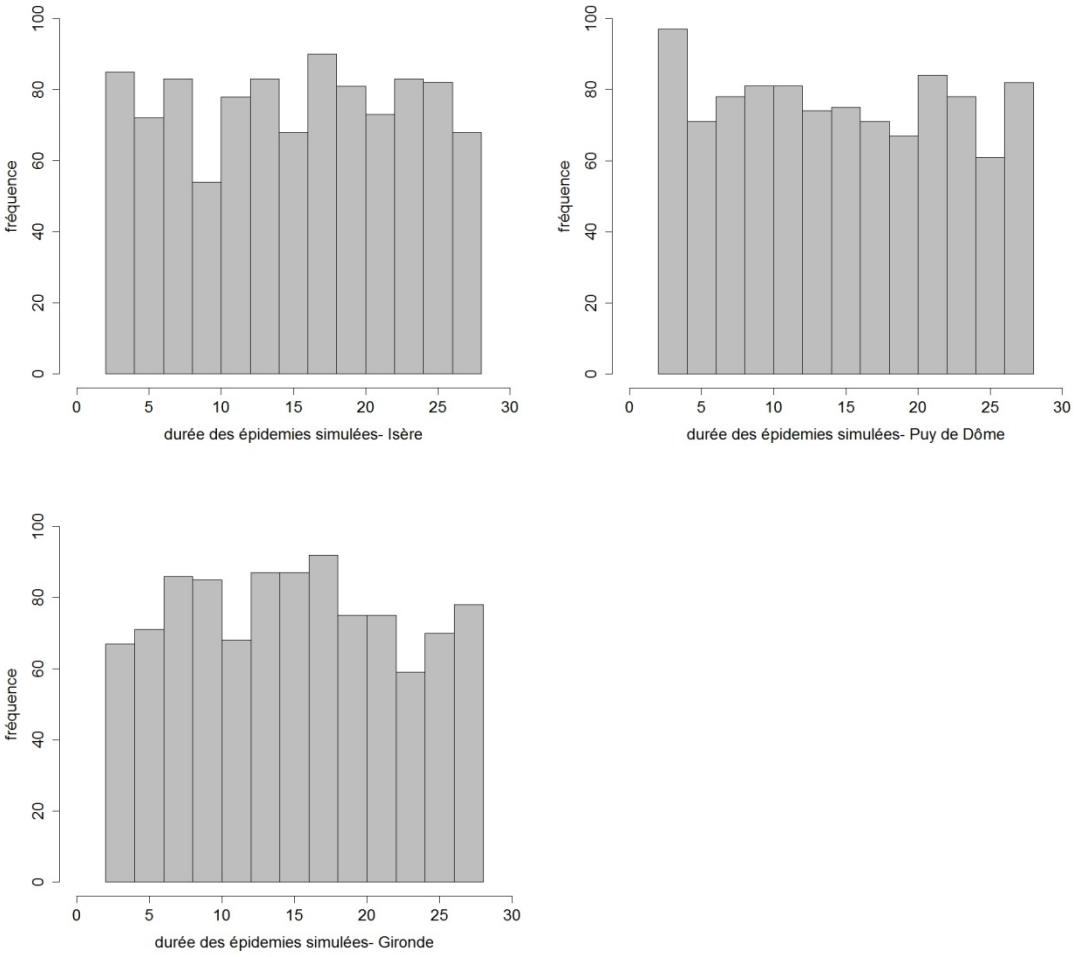
I TABLEAU 2 I

Distribution du nombre de cas simulés (fonction de la variation d'incidence et de la population de l'UDI* qui ont été tirées au sort)

	min	P25	médiane	moyenne	P75	max
Isère	202	400	905	2 151	2 060	154 000
Puy de Dôme	203	367	692	3 860	2 842	95 310
Gironde	223	2 108	4 571	11 000	10 500	179 000

I FIGURE 6 I

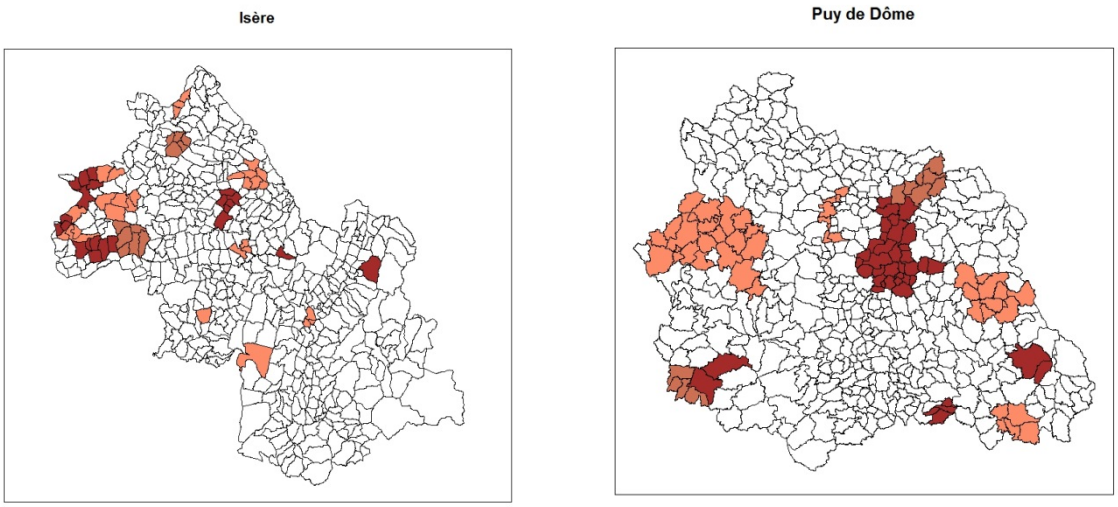
Distribution de la durée des épidémies simulées par département étudié



Des exemples de différentes UDI* utilisées dans les simulations sont présentés dans la Figure 7.

I FIGURE 7 I

Exemples de différentes UDI* utilisées dans les simulations pour les départements de l'Isère et du Puy de Dôme



4.3 Détection avec SaTScan : prise en compte des contours des UDI*

Les résultats de la détection, cluster principal et clusters secondaires, obtenus en choisissant le seuil p <= 0.05 sont présentés.

4.3.1. Calcul de la sensibilité et proportion de fausses alertes

La sensibilité et proportion de fausses alertes sont respectivement : pour l'Isère de 73% (734 épidémies simulées détectées sur 1000 épidémies simulées) et 7% (54 agrégats détectés ne correspondant pas à des épidémies simulées sur 788 agrégats détectés); pour le Puy de Dôme de 73% (728/1 000) et 7% (53/781); pour la Gironde de 90% (904 épidémies simulées détectées sur 1 000 épidémies simulées) et 7% (67 agrégats détectés ne correspondant pas à des épidémies simulées sur 971 agrégats détectées).

Ces indicateurs ont été aussi calculés par classe de population (Tableau 3 et Figure 8) et par nombre de cas (Tableau 4). On peut remarquer qu'à partir de la classe 750-1 000 habitants desservis par l'UDI*, la sensibilité est supérieure à 80% (Figure 8). On n'observe pas de tendance claire quant à la proportion de fausses alertes selon la taille de l'UDI* de l'épidémie simulée.

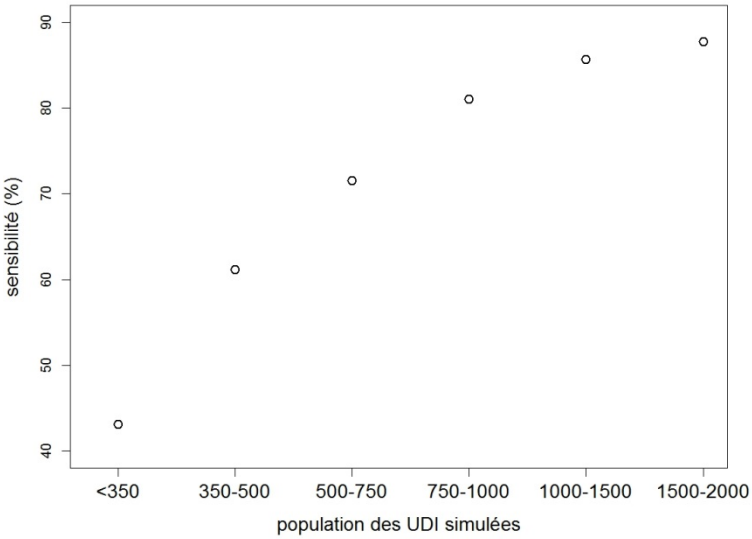
I TABLEAU 3 I

Sensibilité et proportion de fausses alertes en fonction des classes de population des UDI* simulées

Taille UDI*	Sensibilité (nombre de simulations ou d'épidémies simulées)	Proportion de fausses alertes (nombre d'agrégats détectés)
Isère	73% (1000)	7% (788)
200- 500 habitants	46% (321)	13% (169)
500- 1000 habitants	75% (224)	4% (176)
1000- 2000 habitants	85% (193)	6% (174)
2000- 10000 habitants	97% (239)	6% (245)
>=10000 habitants	100% (23)	4% (24)
Puy de Dôme	73% (1000)	7% (781)
200- 500 habitants	52% (385)	6% (216)
500- 1000 habitants	75% (204)	10% (170)
1000- 2000 habitants	84% (128)	4% (112)
2000- 10000 habitants	91% (188)	5% (181)
>=10000 habitants	99% (95)	8% (102)
Gironde	90% (1000)	7% (971)
200- 500 habitants	56% (50)	10% (31)
500- 1000 habitants	75% (63)	10% (52)
1000- 2000 habitants	93% (100)	7% (100)
2000- 10000 habitants	91% (520)	6% (504)
>=10000 habitants	98% (267)	8% (284)
3 départements	79% (3000)	7% (2540)
200- 500 habitants	50% (756)	9% (416)
500- 1000 habitants	75% (491)	7% (398)
1000- 2000 habitants	86% (421)	6% (386)
2000- 10000 habitants	93% (947)	6% (930)
>=10000 habitants	98% (385)	8% (410)

I FIGURE 8 I

La sensibilité en fonction des classes de population de moins de 2 000 habitants



Résultats pour l'Isère, le Puy de Dôme et la Gironde réunis : 1 668 simulations. La première classe correspond à 200-350 habitants (473 simulations), la deuxième classe : 350-500 habitants (283 simulations), la troisième classe : 500-750 (306 simulations), la classe 4 : 750-1 000 (185 simulations), la cinquième classe : 1 000-1 500 (258 simulations) et la dernière classe correspond à 1 500-2 000 habitants (163 simulations).

I TABLEAU 4 I

Sensibilité et proportion de fausses alertes en fonction du nombre de cas simulé

Nombre de cas	Sensibilité (nombre de simulations ou d'épidémies simulées)	Proportion de fausses alertes (nombre d'agrégats détectés)
Isère	73% (1000)	7% (788)
<10 cas	14% (224)	22% (40)
10- 20 cas	71% (222)	9% (172)
>=20 cas	98% (554)	5% (576)
Puy de Dôme	73% (1000)	7% (781)
<10 cas	16% (242)	23% (52)
10- 20 cas	77% (231)	4% (186)
>=20 cas	97% (527)	6% (543)
Gironde	90% (1000)	7% (748)
<10 cas	29% (49)	12% (16)
10- 20 cas	66% (74)	12% (56)
>=20 cas	96% (877)	6% (899)
3 départements	79% (3000)	7% (2540)
<10 cas	16% (515)	21% (108)
10- 20 cas	73% (527)	7% (414)
>=20 cas	97% (1958)	6% (2018)

On remarque que la sensibilité globale est bien plus élevée pour la Gironde (90%) que pour l'Isère ou le Puy de Dôme (73%). Ceci est dû au fait que les UDI* simulées pour la Gironde sont à 89% des UDI* avec plus de 1000 habitants, alors que pour l'Isère et le Puy de Dôme ces UDI* représentent, respectivement, 45% et 41%.

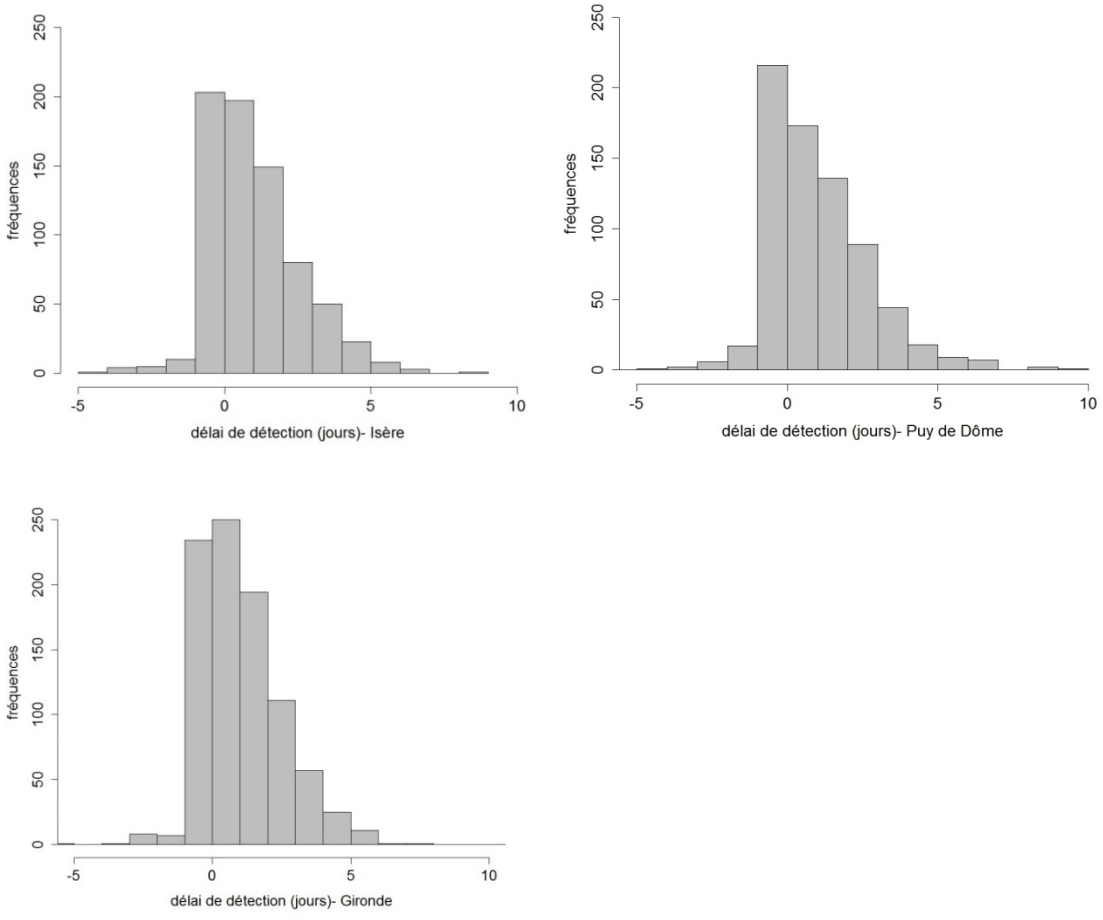
Pour la période hivernale (de décembre à mars), les indicateurs de sensibilité et proportion de fausses alertes sont respectivement : pour l'Isère (nombre de simulations = 307) de 70% et 8%, pour le Puy de Dôme (nombre de simulations = 298) de 67% et 8% et pour la Gironde (nombre de simulations = 321) de 86% et 8%. Pour le reste de l'année (d'avril à novembre), les indicateurs de sensibilité et proportion de fausses alertes sont respectivement : pour l'Isère (nombre de simulations = 693) de 75% et 6%, pour le Puy de Dôme (nombre de simulations = 702) de 75% et 6% et pour la Gironde (nombre de simulations = 679) de 92% et 6%. D'après ces résultats, la méthode de détection est plus performante hors période hivernale (sensibilité un peu plus élevée qu'en hiver et proportion de fausses alertes légèrement plus faible qu'en hiver).

4.3.2. Identification du début de l'épidémie

L'indicateur délai de détection est présenté dans la Figure 9. On observe que 25% des agrégats détectés incluent le premier jour de l'épidémie simulée et 20% manquent au moins les trois premiers jours.

I FIGURE 9 I

Identification du début de l'épidémie



On observe un délai égal à zéro si le premier jour détecté correspond au premier jour simulé. La médiane est à 1 jour (P75 est à 2 jours).

4.3.3. Nombre de communes détectées par rapport au nombre de communes simulées

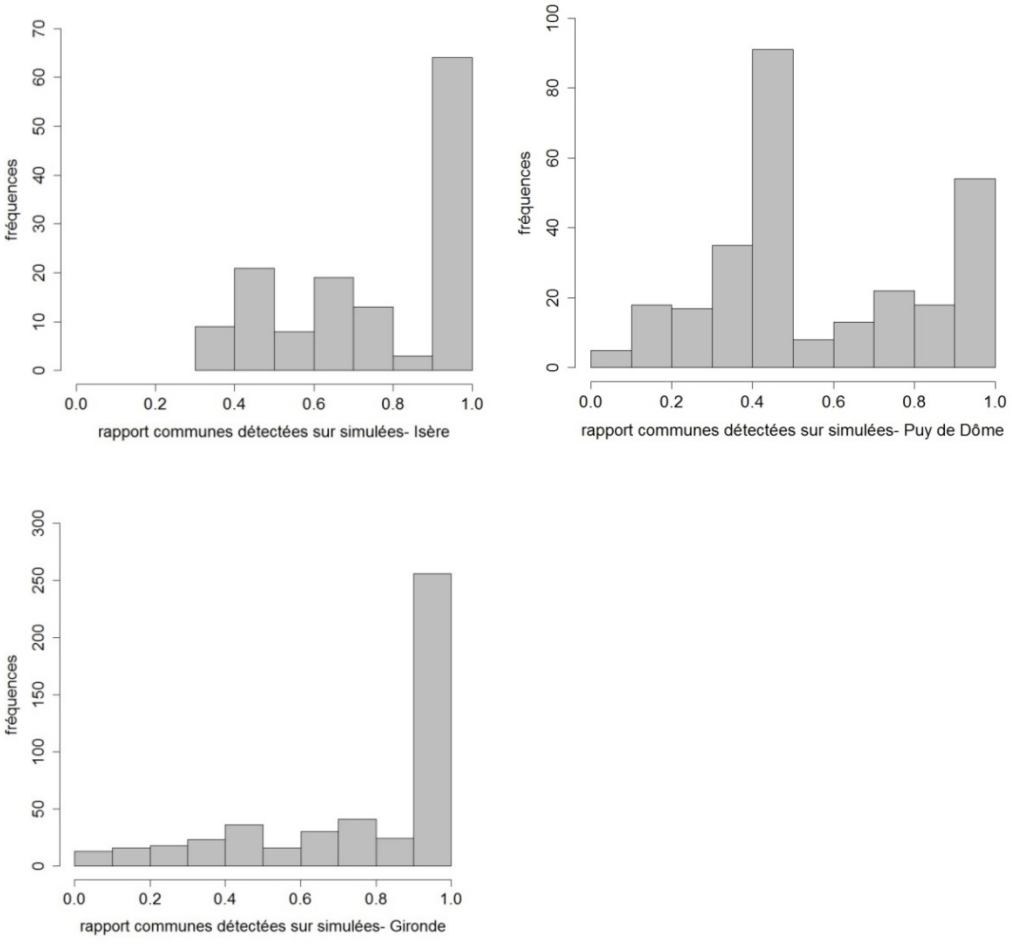
Pour l'Isère, parmi les 1000 épidémies simulées, 162 concernent des UDI* qui desservent plusieurs communes (de 2 à 13 communes). Pour les 137 épidémies correctement détectées (85%), le rapport du nombre de communes détectées sur le nombre de communes concernées par les simulations va de 0.33 à 1 avec la médiane égale à 0.8 (Figure 10).

Pour le Puy de Dôme, parmi les 1 000 épidémies simulées, 372 concernent des UDI* qui desservent plusieurs communes (de 2 à 53 communes). Pour les 281 épidémies correctement détectées (75%), le rapport du nombre de communes détectées sur le nombre de communes simulées va de 0.03 à 1 avec la médiane égale à 0.5 (Figure 10).

Pour la Gironde, parmi les 1 000 épidémies simulées, 521 concernent des UDI* qui desservent plusieurs communes (de 2 à 34 communes). Pour les 473 épidémies correctement détectées (91%), le rapport du nombre de communes détectées sur le nombre de communes simulées va de 0.03 à 1 avec la médiane égale à 1 (Figure 10).

I FIGURE 10 I

Nombre de communes correctement détectées parmi les communes simulées



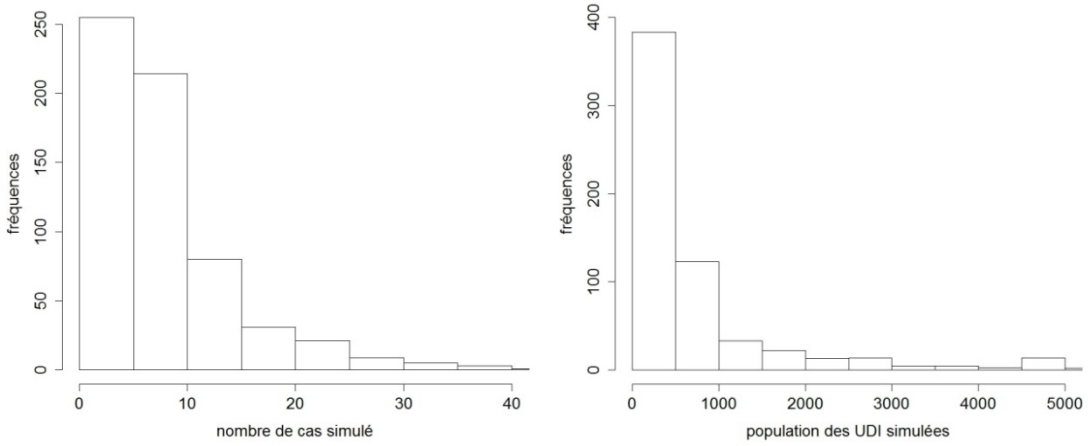
En Isère les UDI* partagées comprennent de 2 à 13 communes (avec une médiane à 3 et P75 à 5) et la population moyenne est de 3 500. Les UDI* partagées dans le Puy de Dôme comprennent de nombreuses communes, de 2 à 53 communes (avec une médiane à 4 et P75 à 8), et elles ont une population importante (population moyenne = 5 200). Les UDI* partagées dans la Gironde comprennent de nombreuses communes, de 2 à 34 communes (avec une médiane à 5 et P75 à 10), et elles ont une population importante (population moyenne = 15 520).

4.3.4. Description des épidémies non détectées

Sur les 1 000 épidémies simulées, 266 (27%) n'ont pas été détectées en Isère, 272 (27%) dans le Puy de Dôme et 96 (10%) en Gironde. Le nombre de cas de ces épidémies simulées non détectées (634 épidémies, 21%) et la population des UDI* correspondantes pour les trois départements montrent qu'il s'agit d'épidémies avec un faible nombre de cas (médiane = 7 et P75 = 11 cas) appartenant à des UDI* desservant un faible nombre d'habitants (médiane = 408 et P75 = 796 habitants) (Figure 11). Parmi les 634 épidémies non détectées, seulement 62 (10%) d'entre elles ont plus de 20 cas. De ces épidémies 39 sur 62 (63%) ont été simulées en hiver (décembre, janvier, février) et 46 sur 62 (75%) ont une durée élevée (de plus de 14 jours).

I FIGURE 11 I

Nombre de cas simulé et population des UDI* correspondantes pour les 634 épidémies simulées qui n'ont pas été détectées



4.3.5. Communes détectées à tort

La proportion de fausses alertes à la commune est égale à 9% (95 communes détectées à tort sur 1085 communes détectées) pour l'Isère, à 3% (53 communes détectées à tort sur 1844 communes détectées) pour le Puy de Dôme et à 2% (67 communes détectées à tort sur 2834 communes détectées) pour la Gironde. Les fausses alertes par nombre de cas détectés (à l'UDI*) sont présentées dans le tableau suivant (Tableau 5). On observe qu'il n'y a pas de règles pour la répartition des fausses alertes selon la taille de l'agrégat, les jours de la semaine ou la période (été/hiver).

I TABLEAU 5 I

Nombre de fausses alertes (exprimé en nombre de communes) en fonction du nombre de cas de l'agrégat détecté

Nombre de cas détectés	Nombre de fausses alertes (%)
Isère	95 (100%)
<10 cas	39 (41%)
10- 20 cas	25 (26%)
>=20 cas	31 (33%)
Puy de Dôme	53 (100%)
<10 cas	18 (34%)
10- 20 cas	25 (47%)
>=20 cas	10 (19%)
Gironde	67 (100%)
<10 cas	20 (30%)
10- 20 cas	23 (34%)
>=20 cas	24 (36%)

Les résultats obtenus avec la matrice de voisins selon la distance sont présentés en Annexe 2. Les indicateurs de sensibilité et spécificité sont très proches de ceux obtenus en prenant en compte les contours des UDI*.

4.4 Détection avec la méthode de comparaison géographique des taux d'incidence

4.4.1. Calcul des indicateurs de sensibilité et proportion de fausses alertes

Les indicateurs de sensibilité et proportion de fausses alertes sont respectivement : pour l'Isère de 71% (714 épidémies simulées détectées sur 1 000 épidémies simulées) et 89% (7 901 agrégats détectés ne correspondant pas à des épidémies simulées sur 8 828 agrégats détectés); pour le Puy de Dôme de 67% (674/1 000) et 89% (15084/ 16958) et pour la Gironde de 86% (859/1 000) et 71% (6 416/9 047).

Ces indicateurs ont été aussi calculés par classe de population de l'UDI* (Tableau 6).

I TABLEAU 6 I

Sensibilité et proportion de fausses alertes en fonction des classes de population des UDI* simulées

Taille de l'UDI*	Sensibilité (nombre de simulations)	Proportion de fausses alertes (nombre d'agrégats communes détectés)
Isère	71% (1000)	89% (8828)
200- 500 habitants	43% (321)	95% (2709)
500- 1000 habitants	73% (224)	91% (1942)
1000- 2000 habitants	86% (193)	89% (1707)
2000- 10000 habitants	94% (239)	82% (2265)
>=10000 habitants	100% (23)	85% (205)
Puy de Dôme	67% (1000)	89% (16958)
200- 500 habitants	42% (385)	97% (6033)
500- 1000 habitants	70% (204)	85% (3217)
1000- 2000 habitants	82% (128)	94% (2075)
2000- 10000 habitants	92% (188)	85% (3286)
>=10000 habitants	96% (95)	60% (2347)
Gironde	86% (1000)	71% (9047)
200- 500 habitants	40% (50)	94% (357)
500- 1000 habitants	68% (63)	89% (441)
1000- 2000 habitants	84% (100)	86% (785)
2000- 10000 habitants	88% (520)	75% (4435)
>=10000 habitants	96% (267)	56% (3029)
3 départements	75% (3000)	84% (29403)
200- 500 habitants	42% (756)	96% (8779)
500- 1000 habitants	71% (491)	93% (5227)
1000- 2000 habitants	84% (421)	91% (4142)
2000- 10000 habitants	90% (947)	80% (7978)
>=10000 habitants	96% (385)	59% (3277)

4.4.2. Description des épidémies non détectées

Sur les 1 000 épidémies simulées, 286 (29%) n'ont pas été détectées en Isère, 326 (33%) n'ont pas été détectées dans le Puy de Dôme et 141 (14%) n'ont pas été détectées en Gironde. On remarque qu'il s'agit d'épidémies avec un faible nombre de cas (médiane = 8 et P75 = 13 cas) appartenant à des UDI* desservant un faible nombre d'habitants (médiane = 419 et P75 = 930 habitants). Parmi les 753 épidémies non détectées, 112 (15%) ont plus de 20 cas. De ces épidémies 75 sur 112 (67%) ont été simulées en hiver (en particulier en décembre, janvier, février) et 73 sur 112 (65%) ont une durée élevée (de plus de 14 jours).

4.4.3. Communes détectées ne correspondant pas à des épidémies simulées

La proportion de fausses alertes à la commune pour l'Isère est égale à 89% (7 901 communes détectées à tort sur 8 828 communes détectées), à 89% (15 086 communes détectées à tort sur 16 958 communes détectées) pour le Puy de Dôme et à 71% (6 416 communes détectées à tort sur 9 047 communes détectées) pour la Gironde (Tableau 7).

I TABLEAU 7 I

Nombre de fausses alertes (nombre de communes) en fonction du nombre de cas de l'agrégat détecté

Nombre de cas simulés	Nombre de fausses alertes (%)
Isère	7901 (100%)
<10 cas	1780 (23%)
10- 20 cas	1770 (22%)
>=20 cas	4351 (55%)
Puy de Dôme	15 086 (100%)
<10 cas	3695 (25%)
10- 20 cas	3469 (23%)
>=20 cas	7922 (53%)
Gironde	6416 (100%)
<10 cas	326 (5%)
10- 20 cas	483 (8%)
>=20 cas	5607 (87%)

4.5 Comparaison des méthodes

Le Tableau 8 reprend les indicateurs de sensibilité et proportion de fausses alertes pour les 3 méthodes testées. Pour la comparaison des méthodes, la proportion de fausses alertes considérée pour les méthodes de Kulldorff est celle calculée à la commune.

I TABLEAU 8 I

Comparaison des méthodes de détection

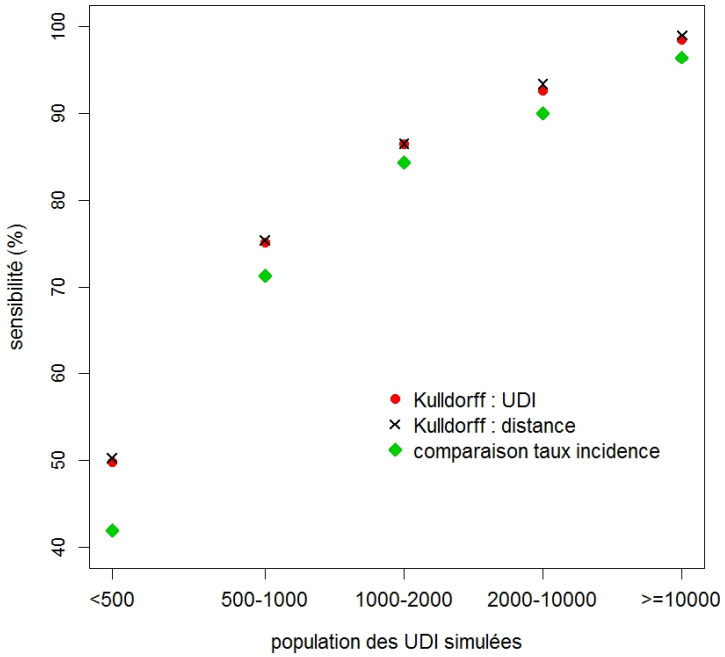
	Sensibilité (nombre de simulations)	Proportion de fausses alertes *à la commune* (nombre d'agrégats- commune détectés)
*Kulldorff avec matrice de voisins UDI**		
Isère	73% (1000)	9% (1085)
Puy de Dôme	73% (1000)	3% (1844)
Gironde	90% (1000)	2% (2834)
3 départements réunis	79% (3000)	4% (5763)
Kulldorff avec matrice de voisins basée sur la distance		
Isère	74% (1000)	11% (1093)
Puy de Dôme	73% (1000)	6% (2106)
Gironde	91% (1000)	6% (3415)
3 départements réunis	79% (3000)	7% (6614)
Méthode de comparaison géographique des taux d'incidence		
Isère	71% (1000)	89% (8828)
Puy de Dôme	67% (1000)	89% (16958)
Gironde	86% (1000)	71% (9047)
3 départements réunis	75% (3000)	84% (34833)

On peut remarquer que la sensibilité des 3 méthodes est assez proche. Par contre la proportion de fausses alertes est de 4% et 7% pour les méthodes de Kulldorff et elle est de 84% pour la méthode de comparaison géographique des taux d'incidence avec 34 833 communes détectées à tort.

La comparaison de ces indicateurs calculés par classe de population est présentée dans les Figures 12 et 13.

I FIGURE 12 I

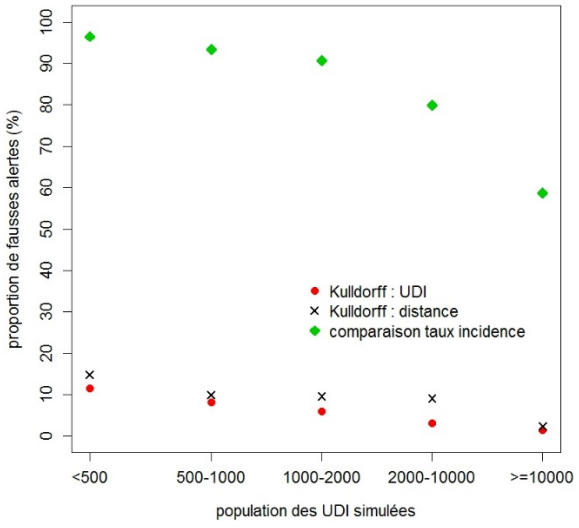
Sensibilité des méthodes en fonction de la population des UDI objet d'une épidémie simulée pour les trois départements réunis



La première classe 200-500 habitants supporte 756 épidémies simulées, la deuxième classe 500-1 000 habitants 491, la troisième classe 1 000-2 000 habitants 421, la quatrième classe 2 000-10 000 habitants avec 947 et la dernière classe correspond à >=10 000 habitants avec 385 épidémies simulées.

I FIGURE 13 I

Proportion de fausses alertes engendrées par les méthodes testées en fonction de la population des UDI objet d'une épidémie simulée pour les trois départements réunis



À part pour les épidémies touchant des UDI* de 200- 500 habitants, la sensibilité de la méthode de comparaison géographique des taux d'incidence est assez proche de celle obtenue avec Kulldorff. Par contre la proportion de fausses alertes est très élevée : la méthode détecte beaucoup d'épidémies et la très grande majorité (89% pour l'Isère et le Puy de Dôme et 71% pour la Gironde) sont des fausses alertes.

4.5.1. Comparaison selon la taille de l'épidémie

La comparaison des méthodes selon la taille de l'épidémie simulée (nombre de cas à détecter) est présentée dans le Tableau 9. Nous avons comparé les méthodes pour des épidémies simulées de plus de 10 cas, de plus de 20 cas, pour des petites épidémies avec 5-10 cas et des épidémies avec 10-20 cas. Le résultat qui nous intéresse particulièrement est celui obtenu pour les épidémies avec >= 10 cas.

I TABLEAU 9 I

Comparaison des méthodes de détection selon la taille des épidémies simulées

	Sensibilité (nombre de simulations)	Proportion de fausses alertes (nombre d'agrégats détectés)	Proportion de fausses alertes *à la commune* (nombre d'agrégats-commune détectés)
Taille de l'épidémie simulée >= 20 cas			
Kulldorff avec matrice de voisins			
Isère	98% (554)	5% (576)	7% (852)
Puy de Dôme	97% (527)	6% (543)	2% (1601)
Gironde	96% (877)	6% (899)	2% (2761)
Kulldorff avec matrice de voisins basée sur la distance			
Isère	98% (554)	4% (572)	8% (850)
Puy de Dôme	96% (527)	6% (539)	5% (1822)
Gironde	97% (877)	5% (899)	6% (3327)
Méthode de comparaison géographique des taux d'incidence			
Isère	96% (554)	-	85% (5096)
Puy de Dôme	95% (527)	-	82% (9618)
Gironde	93% (877)	-	70% (8193)
Taille de l'épidémie simulée >= 10 cas			
Kulldorff avec matrice de voisins			
Isère	91% (776)	6% (748)	8% (1044)
Puy de Dôme	91% (758)	6% (729)	2% (1792)
Gironde	94% (951)	7% (955)	2% (2818)
Kulldorff avec matrice de voisins basée sur la distance			
Isère	91% (776)	5% (744)	9% (1047)
Puy de Dôme	91% (758)	5% (727)	5% (2035)
Gironde	94% (951)	6% (955)	6% (3394)
Méthode de comparaison géographique des taux d'incidence			
Isère	88% (776)	-	87% (7020)
Puy de Dôme	86% (758)	-	86% (13235)
Gironde	89% (951)	-	70% (8713)

Taille de l'épidémie simulée 10- 20 cas			
Kulldorff avec matrice de voisins			
Isère	71% (222)	9% (172)	12% (192)
Puy de Dôme	77% (231)	4% (186)	4% (191)
Gironde	66% (74)	12% (56)	12% (57)
Kulldorff avec matrice de voisins basée sur la distance			
Isère	72% (222)	6% (172)	15% (197)
Puy de Dôme	78% (231)	4% (188)	7% (213)
Gironde	65% (74)	14% (56)	21% (67)
Méthode de comparaison géographique des taux d'incidence			
Isère	69% (222)	-	92% (1924)
Puy de Dôme	63% (231)	-	96% (3617)
Gironde	49% (74)		93% (520)
Taille de l'épidémie simulée 5- 10 cas			
Kulldorff avec matrice de voisins			
Isère	24% (131)	16% (37)	16% (38)
Puy de Dôme	26% (151)	13% (46)	13% (46)
Gironde	36% (36)	13% (15)	13% (15)
Kulldorff avec matrice de voisins basée sur la distance			
Isère	22% (131)	17% (35)	19% (36)
Puy de Dôme	26% (151)	18% (49)	29% (62)
Gironde	42% (36)	12% (17)	15% (20)
Méthode de comparaison géographique des taux d'incidence			
Isère	21% (131)	-	97% (1049)
Puy de Dôme	19% (151)	-	99% (2321)
Gironde	22% (36)	-	97% (250)

À partir de ces résultats, la seule méthode performante est la méthode de détection spatio-temporelle de Kulldorff, indépendamment de la matrice de voisinage retenue. Quand on considère les épidémies avec plus de 10 cas, la sensibilité obtenue avec la méthode de comparaison géographique des taux d'incidence (88% en considérant les trois départements ensemble) est comparable à celle obtenue avec la méthode de Kulldorff (92%), même si toujours plus faible, mais la proportion de fausses alertes est très élevée. La proportion de fausses alertes est en effet de 3% pour la méthode de Kulldorff avec prise en compte des UDI* (6% avec la distance) et elle est de 81% pour la méthode de comparaison géographique des taux d'incidence.

Lors de sa conception, la méthode fondée sur la comparaison de l'incidence n'était pas paramétrée de la même manière que dans cette étude comparative [6]. Les agrégats avec moins de 5 cas par commune et semaine et les risques relatifs inférieurs à 2 (la référence étant l'incidence médiane des communes du département) étaient notamment exclus. Ces contraintes n'ont pas été reprises pour la comparaison des méthodes l'idée étant que les résultats de détection seraient comparés par taille d'agrégat. Le relâchement du paramétrage a conduit à une inflation de fausses alertes.

Reprise du paramétrage de la méthode de comparaison géographique des taux d'incidence
Quand on reprend les conditions d'usage initiales, la proportion de fausses alertes prend des niveaux raisonnables dans certains départements (Isère et Gironde, Tableau 10) mais reste

très élevée dans le Puy-de-Dôme. La sensibilité est par contre plus faible (de 75% à 69%). Compte-tenu du nombre élevé de fausses alertes résiduelles, une étude détaillée des fausses alertes a été effectuée. Elle indique que les fausses alertes se répètent sur un petit nombre de communes, et qu'une même fausse alerte peut apparaitre sur un nombre élevé de simulations.

I TABLEAU 10 I

Indicateurs de sensibilité et proportion de fausses alertes. Résultats obtenus avec la méthode M_incid en ne gardant que les agrégats avec au moins 5 cas et une p-value inférieure à 10^{-9}

	Sensibilité (nombre d'épidémies simulées)	Proportion de fausses alertes (nombre d'agrégats détectés)
Isère	60% (1 000)	4% (797)
Puy de Dôme	56% (1 000)	53% (3 125)
Gironde	79% (1 000)	1% (2 183)
3 départements	69% (3 000)	28% (6 105)

5. DISCUSSION

D'après les critères utilisés pour évaluer la performance et sur la base du schéma de simulation implémenté la méthode de Kulldorff est à préférer à la méthode fondée sur la comparaison des taux d'incidence pour la détection spatio-temporelle des épidémies de GEAm de type hydrique. Le Tableau 11 résume les forces et faiblesses des méthodes testées.

L'intérêt de la méthode de Kulldorff est qu'il s'agit d'une méthode de détection spatio-temporelle. En effet, une différence importante entre la méthode de Kulldorff et la méthode de comparaison de l'incidence réside dans la prise en compte de la composante spatiale. La méthode de Kulldorff est une méthode de détection spatio-temporelle et est appliquée en exploitant les UDI ou les communes limitrophes dans un rayon de 20 km. La méthode de comparaison géographique des taux d'incidence ne prend pas en compte la notion de « voisinage ». Les communes- semaines sont testées les unes après les autres. Chaque commune et chaque semaine sont testées comme si elles étaient indépendantes des autres communes et des autres semaines. Cela influe donc sur les p-values qui sont sous-estimées. Dans la méthode de Kulldorff un seul test est fait.

La variante de Kulldorff fondée sur une matrice de voisinage définie à partir des UDI* présente l'avantage d'inclure indirectement l'hypothèse hydrique *a priori* (la prise en compte du contour des UDI en amont de la recherche de clusters représente un proxy de l'exposition à l'eau du robinet) dans la recherche de cluster de GEAm. Mais elle nécessite une information actualisée sur la correspondance spatiale et populationnelle UDI*- commune. La matrice de voisinage définie à partir des UDI* devrait être préférée dans les départements où les données de population dans la base SISE-Eaux sont renseignées. L'intérêt de cette approche est qu'elle permet de prendre en compte *a priori* la possible distribution géographique des agrégats de cas de GEAm qui ont en commun l'exposition à une même UDI. Les mêmes résultats en termes de sensibilité et en termes de fausses alertes sont obtenus avec la matrice de voisinage définie avec la distance. L'intérêt de cette dernière approche repose dans la simplicité de l'information nécessaire à sa mise en œuvre (aucune information sur les effectifs de populations des UDI dans SISE-Eaux n'est nécessaire).

La détection de clusters de gastro-entérites médicalisées d'origine hydrique repose sur la qualité de l'information géographique disponible et de sa mise à jour. En l'absence d'informations fiables, la matrice de voisinage définie avec la distance peut être considérée.
La fiabilité des données sur la population distribuée par quartier (données SISE-Eaux) devraient évoluer favorablement. Le projet d'une base géographique nationale a pour ambition d'améliorer l'homogénéité et la qualité des données sur l'eau potable, dont le découpage des UDI. Il devrait permettre à terme de disposer d'un découpage fiable des UDI au sein d'une commune partagée entre plusieurs UDI. Le carroyage Insee à 200 m par 200 m permet de calculer la population résidente dans une zone déterminée, la source des données de population étant le revenu fiscal localisé. Si la couche géographique des UDI était de bonne qualité, les effectifs de population desservie par les UDI au sein d'une commune pourraient être évalués sur cette base, au cas où la base SISE-Eaux ne disposerait pas de cette information. Néanmoins, ces données ne suffisent pas à garantir le rattachement d'un cas à une UDI. La liste des clients alimentés par une UDI est détenue par les exploitants du réseau d'eau. En absence de cette information, seule la connaissance de localisation précise du cas, c'est-à-dire de ses coordonnées GPS ou de son adresse postale complète, pourrait garantir le lien entre un cas et une UDI.

I TABLEAU 11 I

Forces et faiblesses des méthodes testées

	Forces	Faiblesses
Kulldorff avec UDI	-Méthode reconnue -Existence d'un logiciel gratuit -Sensibilité élevée -Spécificité élevée -Prise en compte *a priori* du contour des UDI	-Suppose des données de bonne qualité UDI/commune
Kulldorff avec la distance	-Méthode reconnue -Existence d'un logiciel gratuit -Sensibilité élevée -Spécificité élevée -Ne nécessite pas d'autres informations que la distance entre communes	-Pas de prise en compte *a priori* du contour des UDI -Sinon il faut supposer que la distance soit un bon indicateur « d'UDI » ou d'exposition hydrique (adapter la distance à la taille des réseaux)
M_incid	-Bonne sensibilité -Facile à mettre en œuvre	-Mauvaise spécificité -Pas de prise en compte *a priori* de l'origine hydrique et de la dimension spatiale

6. PERSPECTIVES OPÉRATIONNELLES

Cette section présente et discute les conditions pour la mise en œuvre de la surveillance et les objectifs de santé publique en vue de la mise en œuvre en routine du système de détection d'agrégats de cas de GEAm consacré à la surveillance des épidémies d'origine hydrique.

6.1 La sensibilité

Le déploiement des cellules d'intervention en région de Santé publique France (Cire), la rédaction et la diffusion d'un guide d'investigation des épidémies [17], la constitution d'un entrepôt de cas de GEAm à partir de l'exploitation des données de l'Assurance maladie ont permis d'augmenter sensiblement le nombre d'épidémies de GEA d'origine hydrique investiguées depuis 2000. Leur nombre est ainsi passé de 1 à plus de 2 par an en moyenne. Leur détection reposait uniquement sur le signalement d'un excès de cas de GEA par les acteurs de terrain (médecins généralistes, maires, responsables d'institution) ou par l'occurrence d'une contamination fécale de l'eau distribuée. Ces deux sources de signalement ont une sensibilité médiocre. Les médecins généralistes ne voient en moyenne qu'un cas de GEA tous les 2 jours, tandis que l'incidence doit pratiquement décupler pour éveiller la suspicion du médecin quant à la possibilité d'une épidémie locale. De même, les analyses d'eau réglementaires, qui n'ont pas pour objectif la détection des épidémies mais le contrôle de la conformité de la qualité de l'eau, sont peu fréquentes, notamment pour les petites UDI qui sont les plus à risque (entre 1 et 8 analyses par an pour les UDI de moins de 10 000 habitants, selon leur taille). Globalement et bien qu'en hausse, la couverture de la surveillance des épidémies d'origine hydrique reste modeste car limitée par les modalités de leur détection.

Grâce au screening automatisé de la base des données quotidiennes des cas de GEAm au niveau communal, la sensibilité de la détection des épidémies va considérablement augmenter. Une première estimation fondée sur la méthode M_incid [6] indique qu'il pourrait y avoir un millier d'agrégats de plus de 10 cas de GEAm dont l'origine hydrique est possible, par an en France métropolitaine. On peut donc s'attendre à près de 10 agrégats d'au moins 10 cas par département et par an. Par ailleurs, la recherche d'agrégats par la méthode de Kulldorff testée sur 4 départements de la région Auvergne entre 2009 et 2012 a permis d'identifier 11 clusters avec une probabilité forte d'origine hydrique, contre 2 identifiés par le signalement [11]. Ceci suggère une estimation plus basse de l'ordre d'un cas par département et par an. Quel que soit le niveau de sensibilité du screening à l'échelle de la France, il surclassera le niveau actuel.

6.2 La spécificité

Une première estimation pratiquée par l'ARS Normandie situe le travail pour l'enquête environnementale associée à l'investigation d'un agrégat à 1 à 2 jours pour l'investigateur. Suivant les moyens disponibles dans chaque ARS pour cette thématique, il est probable que seules quelques investigations (< 5) soient envisageables pour une année et un département donné.

Le coût associé à l'investigation d'une fausse alerte peut être important et la répétition d'erreur pourrait entrainer la démotivation des acteurs et le discrédit du système de surveillance.

Le manque de sensibilité des méthodes de détection n'est probablement pas le facteur qui limitera l'activité opérationnelle des services de terrain. La sélection de plusieurs agrégats candidats à l'investigation environnementale par année et par département pourrait dépasser dans de nombreux départements la capacité d'investigation de l'autorité sanitaire.

L'intérêt de la recherche de la sensibilité réside aussi dans la possibilité d'estimer l'impact sanitaire global attribuable aux accidents de pollutions fécales des eaux distribuées car les petits agrégats représentent globalement une part importante de l'impact. La prise en considération des petits agrégats permet aussi de mieux évaluer l'impact dû aux très petites UDI (200-500 usagers desservis), voire d'inclure ces UDI dans le projet de prévention vis-à-vis du risque infectieux. Néanmoins, les performances des méthodes étudiées sont moindres dans cette gamme de taille des UDI (sensibilité autour de 50% et proportion de fausses alertes proche de 10 % pour Kulldorff). D'un autre côté, l'intérêt d'évaluer l'impact global des agrégats est lui-même limité par le fait que les accidents ne représentent qu'une part mineure de l'impact total, la part endémique et hyper-endémique étant probablement beaucoup plus élevée que la part épidémique [18].

6.3 L'impact, mesure d'intérêt de santé publique des épidémies

Les moyens engagés dans une investigation dépendent peu de la taille de l'UDI. L'impact sanitaire associé à un agrégat peut être évalué *a priori* par le nombre de cas de GEAm impliqués dans l'agrégat. Cet indicateur d'impact s'impose donc comme le principal indicateur de l'impact de santé publique avec le taux d'attaque.

La priorisation des investigations vers les épidémies de plus fort impact augmente l'efficience de l'effort de prévention à travers d'une part une économie d'échelle et d'autre part par la réduction de la part des investigations inutiles (dirigées vers les fausses alertes). Il est ainsi à la fois prévisible et raisonnable de penser que les investigations excluront les agrégats dont l'impact ne dépasse pas 10 cas.

6.4 Prioriser le repérage des UDI récidivistes

L'analyse des données historiques disponibles de GEAm (2010 et années suivantes) doit permettre d'identifier les installations où les épidémies se répètent. Il est notoire que certaines installations accumulent les épidémies. En Isère, 2 épidémies ont été investiguées à 10 ans d'intervalle sur la même UDI et les mêmes causes ont été identifiées. A plusieurs reprises, l'examen de séries historiques pratiquées au décours d'une épidémie a montré que cette épidémie n'était pas la première [5]. Outre la détection et la prévention des événements, un deuxième objectif est donc de cibler de façon prioritaire les UDI récidivistes pour mettre en œuvre les actions de prévention des risques associés et vérifier l'efficacité des mesures de préventions qui auront été mises en œuvre. Là encore l'impact en nombre de cas cumulés sur l'historique disponible (depuis 2010) peut dicter le degré de priorité des investigations à engager.

Pour cela, la méthode de détection a été adaptée pour prendre en compte les répétitions d'épidémies sur une même UDI*. Des simulations supplémentaires par département ont été réalisées avec plusieurs épidémies simulées sur une même UDI*. Les résultats obtenus avec la méthode du scan spatio-temporel et la matrice des voisins basée sur la distance sont présentés en Annexe 3. Les résultats sont similaires, sensibilité et proportion de fausses alertes, à ceux obtenus avec une seule épidémie simulée par simulation. Mais cette adaptation nécessite d'être plus largement étudiée.

6.5 Prise en compte de la correspondance spatiale UDI-commune

La correspondance spatiale UDI-commune est utilisée pour contraindre la couverture géographique des agrégats détectés à s'inscrire dans le périmètre d'une UDI. Cette approche favorise la sélection d'agrégats attribuables à la contamination de l'eau du robinet et la censure des agrégats dont l'extension ne correspond pas au territoire d'une UDI. Des agrégats trouvés peuvent aussi être circonscrits sur des portions d'UDI, c'est-à-dire des branches de réseau polluées par des retours d'eau en réseau.

L'information nécessaire pour la prise en compte la correspondance UDI-commune est *a minima* la matrice de voisinage indiquant si telle commune est alimentée (au moins en partie) par telle UDI. Une définition plus poussée peut faire intervenir les effectifs de population des intersections [11].

L'option de tenir compte de la correspondance spatiale UDI-commune dans le processus de détection des agrégats nécessite une attention particulière sur les points suivants :

- Les erreurs dans les données SISE-Eaux sont possibles. L'expérience montre qu'elles varient selon les départements. La maintenance d'un système de détection utilisant les données de correspondance UDI-communes devrait donc nécessiter un contrôle préalable (cohérence, données manquantes) ;

- Les changements structuraux des UDI ne sont pas rares. L'interconnexion des réseaux est encouragée pour pallier la pénurie d'eau, en été, quand survient un accident sur la ressource, lors de la maintenance des installations de traitement ou des réservoirs ;

- L'utilisation saisonnière d'interconnexion pour répondre aux afflux de touristes dans certaines régions. Ces changements ne sont pas toujours connus de l'autorité sanitaire. La question de la prise en compte des redistributions saisonnières dans d'éventuelles mises à jour des données devra se poser. Dans ce cas, des erreurs de classements d'exposition sont donc possible en particulier dans les zones touristiques durant les périodes de vacances scolaires.

L'utilisation des données sur la structure des UDI implique leur mise à jour régulière. La faisabilité de la prise en compte de la correspondance spatiale UDI-commune nécessite un échange avec les ARS et être évaluée région par région.

7. CONCLUSION

Les épidémies de gastro-entérites aiguës d'origine hydrique font l'objet d'un programme de surveillance au sein de Santé publique France depuis 1998. Ce travail s'intègre dans le projet de mise en place d'un système de détection automatisé des agrégats de cas de gastro-entérites aiguës médicalisées fournies à partir des données de l'Assurance maladie, liés à la consommation d'eau du robinet. L'objectif de ce système est d'identifier les différents signaux épidémiques et de mettre en place des investigations environnementales complémentaires pour conforter l'origine hydrique. Il doit également apporter un support décisionnel pour formuler des préconisations de prévention dirigées vers les UDI identifiées comme les plus fragiles.

Les performances de deux méthodes (deux variantes de Kulldorff avec des matrices de voisinage différentes et M_incid) ont été étudiées sur 3 000 simulations d'épidémies présentant des profils d'origine hydrique impactant toute une UDI*, sur trois départements français (1 000 simulations par département). Kulldorff présente une bonne sensibilité et une faible proportion de fausses alertes. Son adaptation pour permettre d'identifier des UDI à risque d'épidémie doit être plus largement étudiée. M_incid ne peut pas être utilisée, car cette méthode génère beaucoup trop de fausses alertes.

La méthode de Kulldorff sera appliquée sur des données réelles de plusieurs départements français pilotes, dans le cadre d'un groupe de travail « connexion entre la détection d'agrégats de cas de GEAm et les enquêtes de terrain ». L'origine hydrique des agrégats identifiés par la méthode de Kulldorff avec prise en compte des UDI (ou, selon la qualité des données, la distance) sera évaluée par des acteurs de terrain (binôme (Cire/ARS) pour chaque département pilote), qui pourront diligenter des enquêtes de terrain le cas échéant. Un des objectifs de ce groupe de travail est de définir les conditions de mise en œuvre de la méthode de détection, et les critères à retenir, en vue de fournir une information adaptée à la surveillance en routine et à la prévention des épidémies de GEA portée par l'eau du robinet.

Références bibliographiques

[1] Bounoure F, Beaudeau P, Mouly D, Skiba M, Lahiani-Skiba M. Syndromic surveillance of acute gastroenteritis based on drug consumption. Epidemiol Infect. 2011;139(9):1388-95.

[2] Beaudeau P, Bentayeb M, Corso M, Rambaud L, Galey C. Les données de l'entrepôt de cas de gastro-entérite médicalisés issues du Sniiram : description, qualité et utilisation. Saint-Maurice : Santé publique France; 2017. 40 p. Disponible: www.santepubliquefrance.fr

[3] Beaudeau P, Bounoure F. Évaluation épidémiologique d'indicateurs d'incidence des gastroentérites fondés sur les données de l'Assurance maladie. Environnement, Risques & Santé. 2006;5(5):373-82.

[4] Beaudeau P. Surveillance syndromique des gastro-entérites aigües : une opportunité pour la prévention du risque infectieux attribuable à l'ingestion d'eau du robinet. Rennes: Université de Rennes 1; 2012. 244 p.

[5] Mouly D, Van Cauteren D, Vincent N, Vaissiere E, Beaudeau P, Ducrot C, *et al.* Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases. Epidemiol Infect. 2016;144(3):591-601.

[6] Rambaud L, Galey C, Beaudeau P. Automated detection of case clusters of waterborne acute gastroenteritis from health insurance data-pilot study in three French districts. Journal of Water and Health 2015.

[7] Kulldorff M, Heffernan R, Hartman J, Assuncao R, Mostashari F. A space-time permutation scan statistic for disease outbreak detection. PLoS Med. 2005;2(3):e59.

[8] Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods. 1997;26.

[9] Texier G, Gaudart J, Queyriaux B. Techniques d'analyse spatiale. In: Astagneau P, Ancelle T, (dir.).Surveillance épidémiologique: Principes, méthodes et applications en santé publique. Paris : Editions Lavoisier; 2011. p. 57-66.

[10] Kulldorff M, Information Management Services Inc. SaTScan v9.1 : Sofware for the spatial, temporal and space-time statistics. 2015. [consulté le 19/02/2016]. Disponible: http://www.satscan.org/

[11] Coly S, Vincent N, Vaissière E, Charras-Garrido M, Gallay A, Ducrot C, *et al.* Detection of waterborne disease outbreaks: an integrated approach using health administrative databases. Journal of Water and Health Under review.

[12] Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. Stat Med. 2013;32(7):1206-22.

[13] Tuppin P, de Roquefeuil L, Weill A, Ricordeau P, Merliere Y. French national health insurance information system and the permanent beneficiaries sample. Rev Epidemiol Sante Publique. 2010;58(4):286-90.

[14] Majowicz SE, Hall G, Scallan E, Adak GK, Gauci C, Jones TF, *et al.* A common, symptom-based case definition for gastroenteritis. Epidemiol Infect. 2008;136(7):886-94.

[15] Beaudeau P, de Valk H, Vaillant V, Mannschott C, Tillier C, Mouly D, *et al.* Lessons learned from ten investigations of waterborne gastroenteritis outbreaks, France, 1998-2006. Journal of Water and Health. 2008;6(4):491-503.

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

[17] Galey C. Guide d'investigation des épidémies d'infections liées à l'ingestion d'eau de distribution. Deuxième édition. Saint-Maurice : Santé publique France; 2017. 60 p. Disponible: www.santepubliquefrance.fr

[18] Beaudeau P. Évaluation et caractérisation du risque d'origine fécale véhiculé par l'eau de distribution en France. État des lieux et perspectives en matière de recherche et de surveillance. Saint-Maurice : Institut de veille sanitaire; 2016. 26 p. Disponible: http://www.invs.sante.fr

[19] Zhang Z, Assunção R, Kulldorff M. Spatial Scan Statistics Adjusted for Multiple Clusters. Journal of Probability and Statistics. 2010;2010:11.

ANNEXES

Annexe 1. Description des départements étudiés

Isère

Le département de l'Isère a 1 224 993 habitants (Recensement 2012, Insee) et il est composé de 533 communes avec une population moyenne de 2 246 habitants (minimum de 12 habitants et maximum de 155 600 habitants). 25% des communes de l'Isère ont moins de 434 habitants et 25% des communes ont plus de 1 910 habitants.

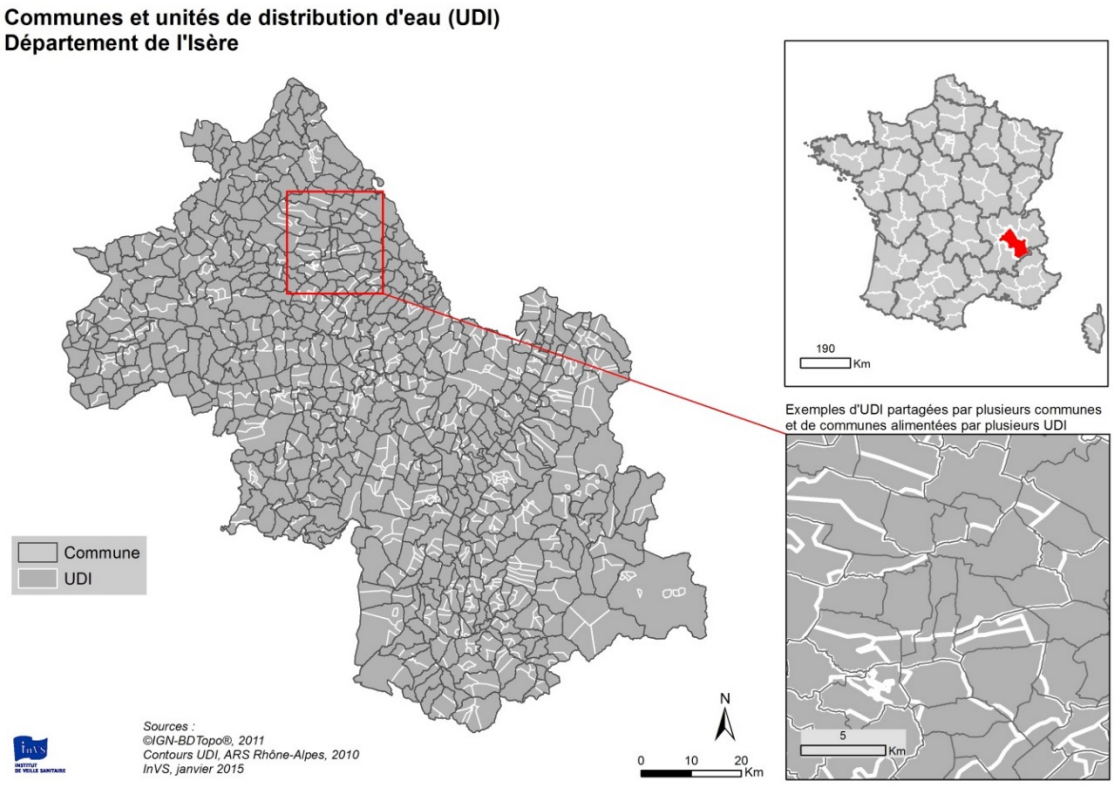
En Isère, il y a 864 UDI dont 779 non partagées (complètement inclues dans une commune) et 85 sont partagées (ce qui représente 335 000 habitants, 28% de la population totale) (base SISE-Eaux). La population desservie par les UDI varie de 1 à 154 000 habitants avec une moyenne de 1 384 habitants. 25% des UDI ont moins de 50 habitants et 25% ont plus de 982 habitants.

Pour les simulations nous avons retenu les 457 UDI de plus de 200 habitants. Ces UDI représentent 98% de la population du département. La population varie alors de 200 à 154 000 avec une moyenne de 2 562 habitants desservis. Les 85 UDI partagées desservent de 2 à 13 communes (médiane = 3 et P75 = 4).

La Figure 14 présente les UDI et les communes de l'Isère. La cartographie et les suivantes ont été réalisées à partir des bases de données géographiques gérées par les ARS, lesquelles présentent des écarts plus ou moins importants avec les informations de la base SISE-Eaux.

I FIGURE 14 I

Communes et UDI de l'Isère



Puy de Dôme

Le département du Puy de Dôme a 638 092 habitants (Recensement 2012, INSEE) et il est composé de 470 communes avec une population moyenne de 1 333 habitants (minimum de 22 habitants et maximum de 139 500 habitants). 25% des communes du Puy de Dôme ont moins de 223 habitants et 25% ont plus de 907 habitants.

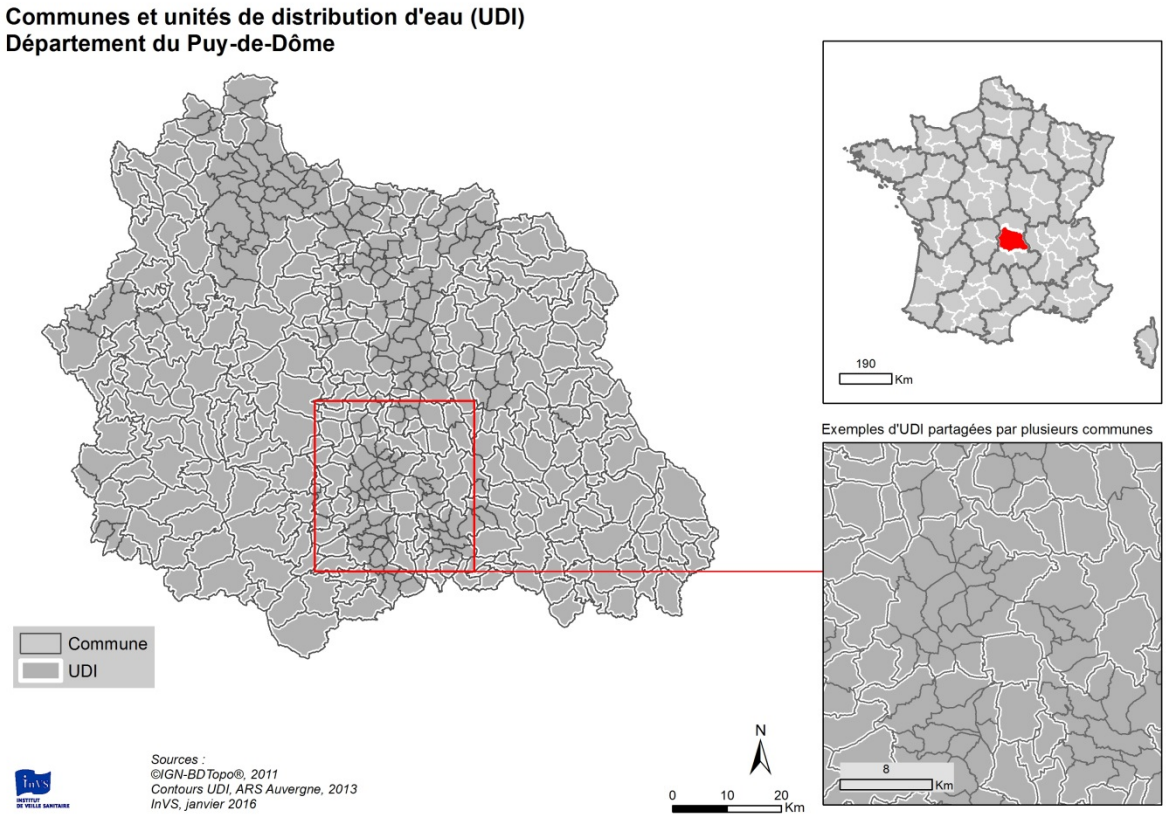
Dans le Puy de Dôme il y a 676 UDI dont 589 sont non partagées et 87 sont partagées (ce qui représente 300 000 habitants, 48% de la population totale) (base SISE-Eaux). La population varie de 1 à 95 310 habitants avec une moyenne de 920 habitants desservis. 25% des UDI ont moins de 13 habitants et 25% ont plus de 200 habitants.

Pour les simulations nous avons retenu les 172 UDI avec plus de 200 habitants. Ces UDI représentent 97% de la population du département. La population varie alors de 200 à 95 310 avec une moyenne de 3 534 habitants desservis. Les 66 UDI partagées desservent de 2 à 67 communes (médiane = 4 et P75 = 7).

La Figure 15 présente les UDI et les communes du Puy de Dôme.

I FIGURE 15 I

Communes et UDI du Puy de Dôme



Gironde

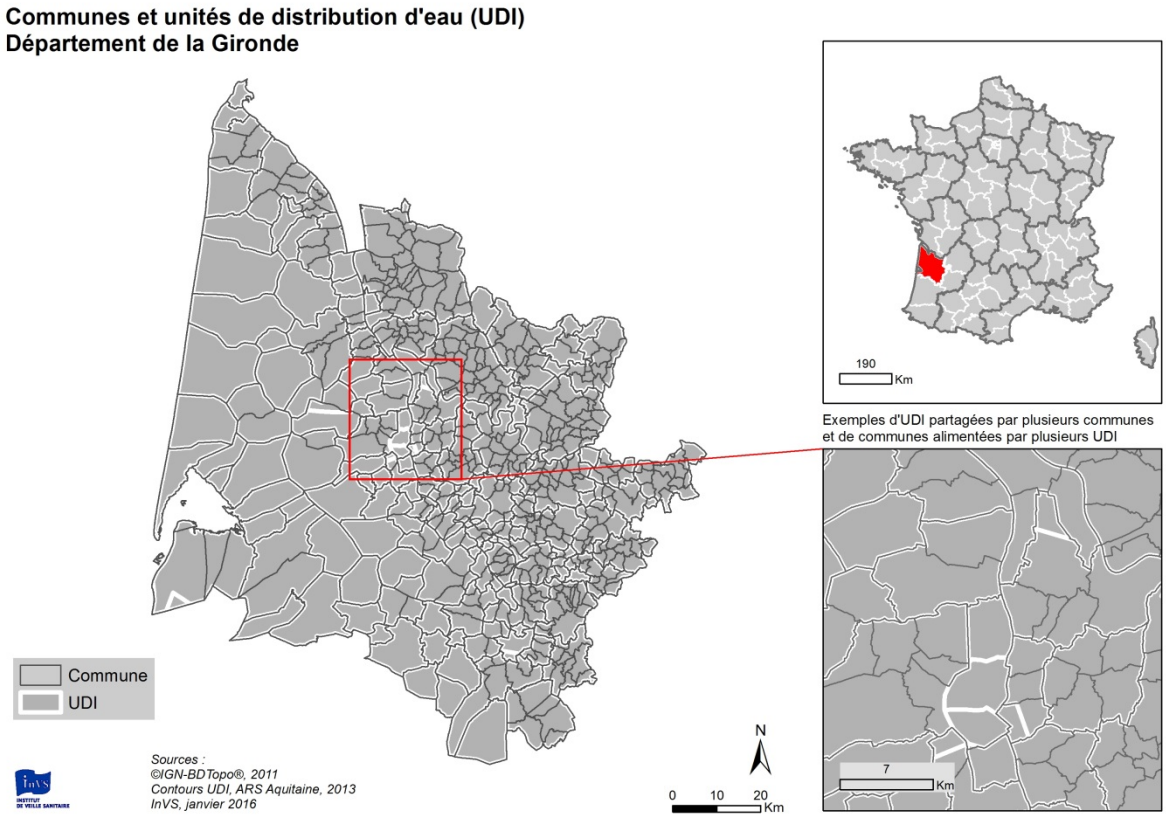
Le département de la Gironde a 1 483 712 habitants (Recensement 2012, INSEE) et il est composé de 542 communes avec une population moyenne de 2 658 habitants (minimum de 47 habitants et maximum de 237 700 habitants). 25% des communes de la Gironde ont moins de 310 habitants et 25% ont plus de 1 826 habitants.

Dans la Gironde il y a 130 UDI dont 62 sont non partagées et 68 sont partagées (ce qui représente 985 882 habitants, 69% de la population totale) (base SISE-Eaux). La population varie de 88 à 179 000 habitants avec une moyenne de 10 930 habitants desservis. Un 25% des UDI a moins de 2 234 habitants et un 25% a plus de 9 726 habitants.

Pour les simulations nous avons retenu les 128 UDI avec plus de 200 habitants. Ces UDI représentent 99.98% de la population du département. La population varie alors de 223 à 179 000 avec une moyenne de 11 100 habitants desservis. Les 68 UDI partagées desservent de 2 à 34 communes (médiane = 5 et P75 = 10).
La Figure 16 présente les UDI et les communes de la Gironde.

I FIGURE 16 I

Communes et UDI de la Gironde



Annexe 2. Kulldorff avec matrice de voisins selon la distance

Calcul des indicateurs de sensibilité et proportion de fausses alertes

Les indicateurs de sensibilité et proportion de fausses alertes sont respectivement : pour l'Isère de 74% (736 épidémies simulées détectées sur 1 000 épidémies simulées) et 6% (48 agrégats détectés ne correspondant pas à des épidémies simulées sur 784 agrégats détectés); pour le Puy de Dôme de 73% (728/1 000) et 7% (53/781) et pour la Gironde de 91% (915/ 1000) et 6% (58/ 973).

Ces indicateurs ont été aussi calculés par classe de population (Tableau 12) et par nombre de cas (Tableau 13).

I TABLEAU 12 I

Sensibilité et proportion de fausses alertes en fonction des classes de population des UDI* simulées

Taille de l'UDI*	Sensibilité (nombre de simulations ou d'épidémies simulées)	Proportion de fausses alertes (nombre d'agrégats détectés)
Isère	74% (1000)	6% (784)
200- 500 habitants	46% (321)	11% (166)
500- 1000 habitants	76% (224)	3% (176)
1000- 2000 habitants	85% (193)	6% (175)
2000- 10000 habitants	97% (239)	5% (244)
>=10000 habitants	100% (23)	0% (23)
Puy de Dôme	73% (1000)	7% (781)
200- 500 habitants	53% (385)	8% (221)
500- 1000 habitants	75% (204)	8% (167)
1000- 2000 habitants	84% (128)	8% (116)
2000- 10000 habitants	91% (188)	4% (178)
>=10000 habitants	99% (95)	5% (99)
Gironde	91% (1000)	6% (973)
200- 500 habitants	60% (50)	9% (33)
500- 1000 habitants	75% (63)	8% (51)
1000- 2000 habitants	92% (100)	6% (98)
2000- 10000 habitants	93% (520)	7% (518)
>=10000 habitants	99% (267)	3% (273)
3 départements	79% (3000)	6% (2538)
200- 500 habitants	50% (756)	9% (420)
500- 1000 habitants	75% (491)	6% (394)
1000- 2000 habitants	86% (421)	6% (389)
2000- 10000 habitants	93% (947)	6% (940)
>=10000 habitants	99% (385)	3% (395)

I TABLEAU 13 I

Sensibilité et proportion de fausses alertes en fonction du nombre de cas simulé

Nombre de cas	Sensibilité (nombre de simulations ou d'épidémies simulées)	Proportion de fausses alertes (nombre d'agrégats détectés)
Isère	74% (1000)	6% (784)
<10 cas	13% (224)	27% (40)
10- 20 cas	72% (222)	6% (172)
>=20 cas	98% (554)	4% (572)
Puy de Dôme	73% (1000)	7% (781)
<10 cas	16% (242)	26% (54)
10- 20 cas	78% (231)	4% (188)
>=20 cas	96% (527)	6% (539)
Gironde	91% (1000)	6% (973)
<10 cas	33% (49)	11% (18)
10- 20 cas	65% (74)	14% (56)
>=20 cas	97% (877)	5% (899)
3 départements	79% (3000)	6% (2538)
<10 cas	16% (515)	24% (112)
10- 20 cas	74% (527)	6% (416)
>=20 cas	97% (1958)	5% (2010)

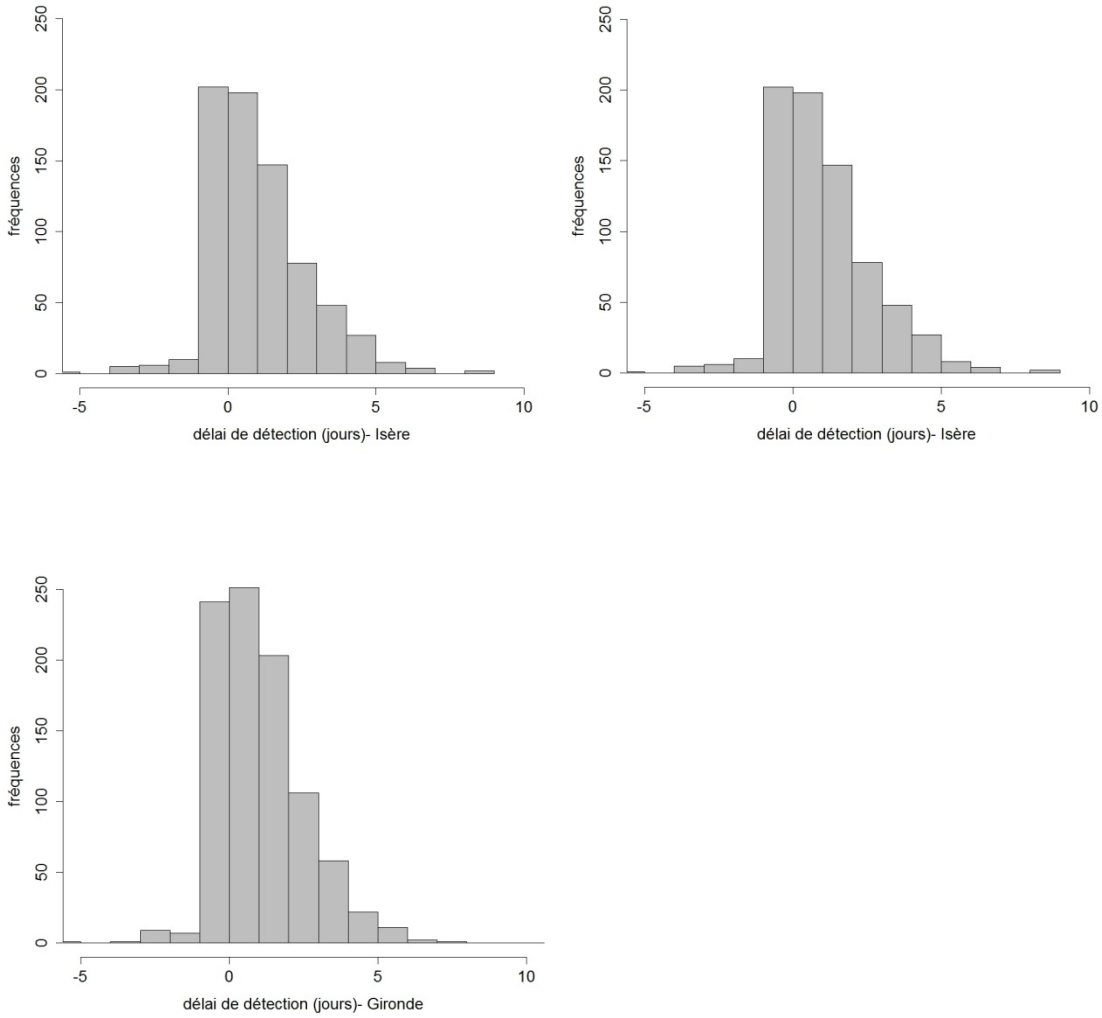
Pour la période hivernale (de décembre à mars), les indicateurs de sensibilité et proportion de fausses alertes pour l'Isère (nombre de simulations = 307) sont respectivement de 70% et 4%, pour le Puy de Dôme (nombre de simulations = 298) de 68% et 8% et pour la Gironde (nombre de simulations = 321) de 88% et 7%. Pour le reste de l'année (d'avril à novembre), les indicateurs de sensibilité et proportion de fausses alertes pour l'Isère (nombre de simulations = 693) sont de 75% et 7%, pour le Puy de Dôme (nombre de simulations = 702) de 75% et 6% et pour la Gironde (nombre de simulations = 679) de 93% et 6%. La sensibilité est un peu plus faible l'hiver.

Identification du début de l’épidémie

L’indicateur délai de détection est présenté dans la Figure 17.

I FIGURE 17 I

Identification du début de l’épidémie. La médiane est à 1 jour (P75 est à 2 jours)



Nombre de communes détectées par rapport au nombre de communes simulées

Pour l'Isère, parmi les 1 000 épidémies simulées, 162 concernent des UDI* qui desservent plusieurs communes.

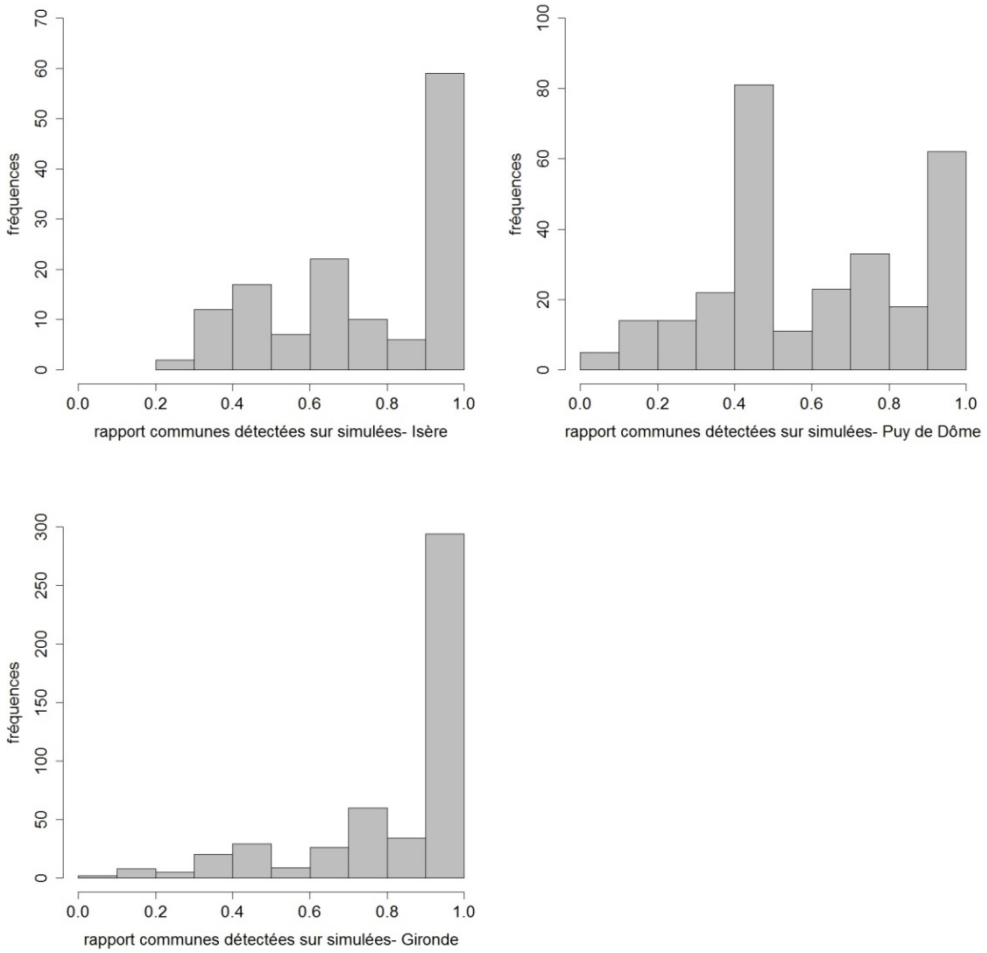
Pour les 135 épidémies correctement détectées (83%), le rapport du nombre de communes détectées sur le nombre de communes simulées va de 0.2 à 1 avec la médiane égale à 0.8 (Figure 18).

Pour le Puy de Dôme, parmi les 1000 épidémies simulées, 372 concernent des UDI* qui desservent plusieurs communes. Pour les 282 épidémies correctement détectées (76%), le rapport du nombre de communes détectées sur le nombre de communes simulées va de 0.03 à 1 avec la médiane égale à 0.6 (Figure 18).

Pour la Gironde, en utilisant la distance pour définir la matrice de voisinage, 487 épidémies parmi les 521 concernant des UDI* desservant plusieurs communes sont correctement détectées (93%) et le rapport du nombre de communes détectées sur le nombre de communes simulées va de 0.03 à 1 avec la médiane égale à 1.

I FIGURE 18 I

Nombre de communes correctement détectées parmi les communes simulées

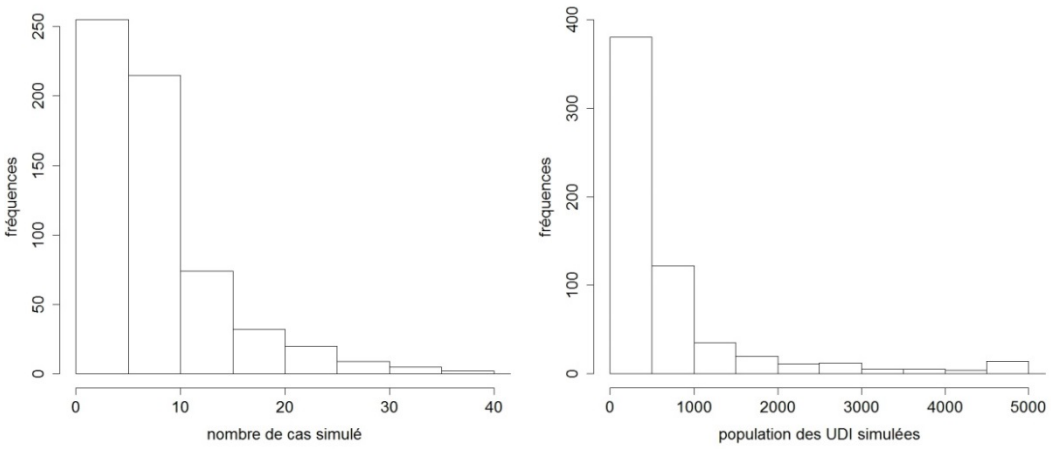


Description des épidémies non détectées

Sur les 1 000 épidémies simulées, 264 (26%) n'ont pas été détectées en Isère, 272 (27%) n'ont pas été détectées dans le Puy de Dôme et 85 (8%) n'ont pas été détectées en Gironde. La Figure 19 présente le nombre de cas de ces épidémies simulées non détectées (621, 21%) et la population des UDI* correspondantes pour les trois départements étudiés. On remarque qu'il s'agit d'épidémies avec un faible nombre de cas (médiane = 6 et P75 = 10 cas) appartenant à des UDI* desservant un faible nombre d'habitants (médiane = 400 et P75 = 756 habitants). Parmi les 621 épidémies non détectées, 62 (10%) ont plus de 20 cas. De ces épidémies 39 sur 62 (63%) ont été simulées en hiver (en particulier en décembre, janvier, février) et 46 sur 62 (75%) ont une durée élevée (de plus de 14 jours).

I FIGURE 19 I

Nombre de cas simulés et population des UDI* correspondantes des 621 épidémies simulées qui n'ont pas été détectées



Communes détectées ne correspondant pas à des épidémies simulées

La proportion de fausses alertes à la commune pour l'Isère est égale à 11% (116 communes détectées à tort sur 1 093 communes détectées), à 6% (136 communes détectées à tort sur 2 106 communes détectées) pour le Puy de Dôme et à 6% (208 communes détectées à tort sur 3 415 communes détectées) pour la Gironde. Comme attendu, la proportion de fausses alertes calculée à la commune, 11% pour l'Isère, 6% pour le Puy de Dôme et 6% pour la Gironde, est plus élevée qu'en utilisant la matrice des voisins définie à partir des UDI*, 9% pour l'Isère, 3% pour le Puy de Dôme et 2% pour la Gironde. Les fausses alertes par nombre de cas détectés sont présentées dans le tableau suivant (Tableau 14). On observe qu'il n'y a pas de règles pour la répartition des fausses alertes selon la taille de l'agrégat, les jours de la semaine ou la période (été/hiver).

I TABLEAU 14 I

Nombre de fausses alertes (nombre de communes) en fonction du nombre de cas de l'agrégat détecté

Nombre de cas détectés	Nombre de fausses alertes (%)
Isère	116 (100%)
<10 cas	41 (35%)
10- 20 cas	33 (29%)
>=20 cas	42 (36%)
Puy de Dôme	136 (100%)
<10 cas	73 (54%)
10- 20 cas	38 (28%)
>=20 cas	25 (18%)
Gironde	208 (100%)
<10 cas	113 (54%)
10- 20 cas	34 (16%)
>=20 cas	61 (30%)

Annexe 3. Répétitions d'épidémies

La méthode de détection a été adaptée pour prendre en compte les répétitions d'épidémies sur une même UDI*. Le cluster principal est enlevé du jeu de données (communes détectées sur la période détectée) et la détection est relancée sur le « nouveau » jeu de données [19]. Des simulations supplémentaires par département ont été réalisées avec 2 à 20 épidémies simulées sur une même UDI*. Les résultats obtenus avec la méthode du scan spatio-temporel et la matrice des voisins basée sur la distance sont présentés dans le Tableau 15. Les résultats sont similaires, sensibilité et proportion de fausses alertes, à ceux obtenus avec une seule épidémie simulée par simulation (Tableau 12).

I TABLEAU 15 I

Répétition d'épidémies simulées dans une même UDI*: sensibilité et proportion de fausses alertes de la méthode Kulldorff basée sur la distance

	Sensibilité (nombre d'épidémies simulées)	Proportion de fausses alertes (nombre d'agrégats détectés)
Isère	76% (559)	5% (449)
Puy de Dôme	72% (626)	1% (453)
Gironde	91% (604)	9% (604)