

*Santé environnement*

# Détection automatisée d'agrégats de cas de gastro-entérites aiguës dans trois départements, France, 2009-2012

Utilisation pour la prévention des épidémies  
d'origine hydrique

Loïc Rambaud, Catherine Galey, Agnès Guillet, Magali Corso,  
Dieter Van Cauteren, Pascal Beaudeau

## Sommaire

Abréviations	2
<b>1. Introduction</b>	<b>3</b>
<b>2. Matériel et méthodes</b>	<b>5</b>
2.1 Données sources	5
2.2 Préparation des données	5
2.3 Détection d'agrégats de cas de GEAm	6
2.4 Détermination du nombre de cas attendus	7
2.5 Test du nombre de cas observés contre le nombre de cas attendus	9
2.6 Présentation des résultats	10
2.7 Évaluation de la sensibilité des méthodes par détection d'agrégats simulés	10
<b>3. Résultats</b>	<b>12</b>
3.1 Fréquence et durée des agrégats	12
3.2 Communes à risque chronique	14
3.3 Nombre de cas impliqués	14
3.4 Variations saisonnières de la détection	16
3.5 Évaluation de la sensibilité des méthodes	16
3.6 Impact national attendu	17
<b>4. Discussion</b>	<b>18</b>
4.1 Comparaison des deux méthodes de détection A et B	18
4.2 Imputation des agrégats à l'eau : une question ouverte	20
4.3 Information utile aux investigateurs de terrain	23
<b>5. Conclusion</b>	<b>26</b>
Références bibliographiques	27
Annexes	29

# **Détection automatisée d'agrégats de cas de gastro-entérites aiguës dans trois départements, France, 2009-2012**

Utilisation pour la prévention des épidémies d'origine hydrique

## **Ont participé à ce rapport**

Loïc Rambaud, Catherine Galey, Agnès Guillet, Magali Corso, Dieter Van Cauteren et Pascal Beaudeau –  
Département santé environnement (DSE), Institut de veille sanitaire (InVS)

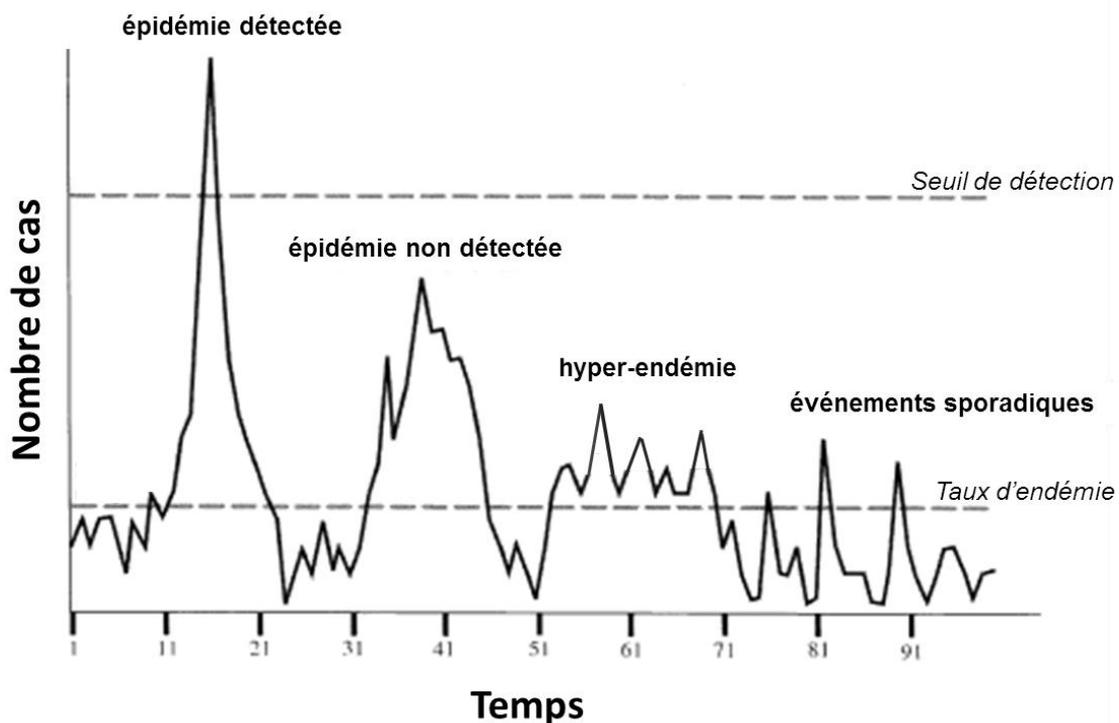
## Abréviations

AC	Agrégat consolidé
AH	Agrégat hebdomadaire
ARS	Agence régionale de santé
Cire	Cellule de l'InVS en région
CxS	Commune x semaine
DO	Déclaration obligatoire
ES	Épidémie simulée
GEA	Gastro-entérite aiguë
GEAm	Gastro-entérite aiguë médicalisée
NCO	Nombre de cas de GEAm observé
NCA	Nombre de cas de GEAm attendu
RIMref	Ratio d'incidence médian du secteur de référence
RR	Risque relatif
Sniir-AM	Système national d'informations inter régimes de l'Assurance Maladie
Tiac	Toxi-infection alimentaire collective (excluant l'eau)
TIMref	Taux d'incidence médian du secteur de référence
UD	Unité de distribution

# 1 Introduction

La surveillance du risque d'origine fécale lié à l'eau du robinet se focalise traditionnellement sur les épidémies [1,2], même si le risque endémique est plus important [3]. Depuis le début des années 2000, l'Institut de veille sanitaire (InVS) recense ainsi annuellement en France en moyenne une à deux épidémies de gastro-entérites aiguës (GEA) d'origine hydrique, sur la base des investigations menées par les Agences régionales de santé (ARS) et les Cellules de l'InVS en région (Cire) [1]. La détection de ces épidémies repose sur le signalement, par les acteurs de terrain, d'agrégats de cas de GEA ou d'analyses microbiologiques non-conformes de l'eau distribuée réalisées dans le cadre du contrôle réglementaire. Ce mode de surveillance se focalise sur les agrégats impliquant de nombreux cas et n'atteint vraisemblablement pas l'exhaustivité. Certains agrégats, de faible ampleur et donc difficilement identifiables, passent encore probablement inaperçus. L'amélioration de ce système de surveillance passe par une augmentation des capacités de détection d'agrégats de GEA d'origine hydrique. Ceci permettrait dans un premier temps de mieux caractériser l'étendu du risque de GEA d'origine hydrique (figure 1) et d'espérer un important gain de santé publique en France.

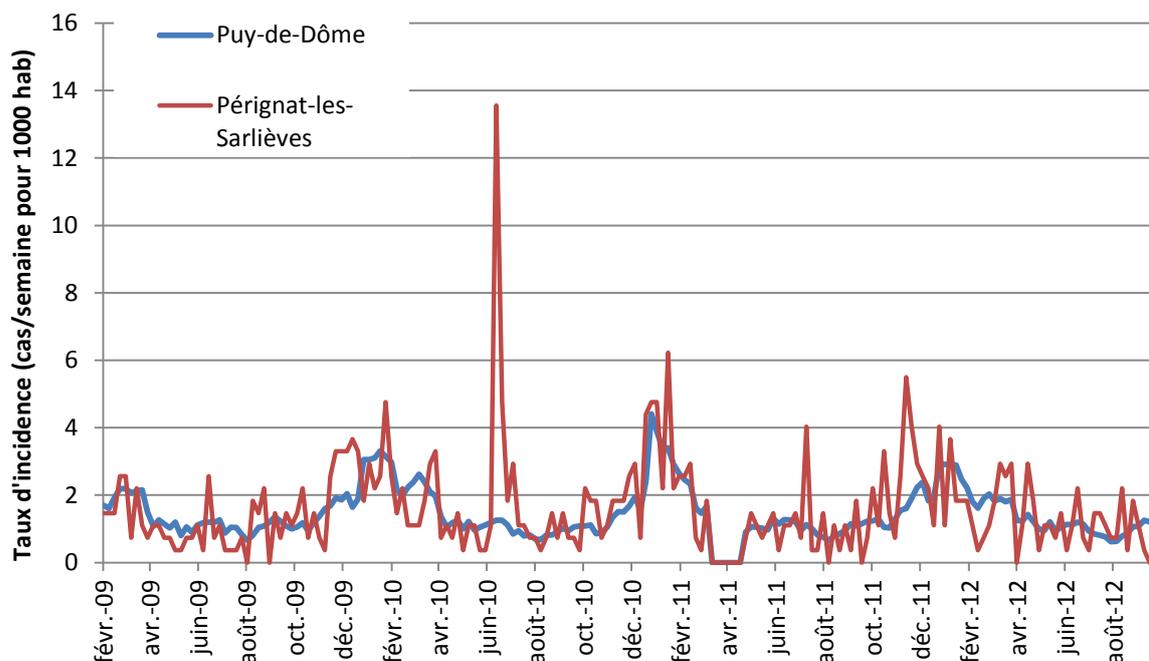
**Figure 1 : Limites du système actuel de surveillance du risque de gastro-entérites aiguës d'origine hydrique en France**



Le développement de bases de données médico-tarifaires offre des possibilités nouvelles. L'exploitation de la base de données du Système national inter régimes de l'Assurance Maladie (Sniir-AM) [4] permet ainsi de disposer d'une estimation du nombre de cas quotidiens de GEA médicalisées (GEAm) à l'échelle de la commune sur l'ensemble de la France [5].

Aujourd'hui, cet indicateur est utilisé par l'InVS pour confirmer rétrospectivement l'existence des agrégats suspectés en comparant une zone cible à une zone de référence, comme par exemple pour l'épidémie de GEA survenue en juillet 2010 sur la commune de Pérignat-les-Sarlièves (figure 2), comparée au département du Puy-de-Dôme. Le cas échéant, les données issues de l'exploitation du Sniir-AM permettent de caractériser ces agrégats en déterminant leur impact sanitaire, leur durée et leurs éventuelles répétitions dans le temps.

**Figure 2 : Taux d'incidence des GEAm dans le département du Puy-de-Dôme et sur la commune de Pérignat-les-Sarlièves (du 1<sup>er</sup> février 2009 au 31 août 2012)**



La détection automatisée prospective d'agrégats de cas de GEAm devrait améliorer le système de surveillance, en permettant :

- 1) d'identifier un plus grand nombre d'agrégats ;
- 2) d'identifier les récurrences et les communes les plus à risque ;
- 3) d'informer les autorités sanitaires locales de santé de façon ciblée afin qu'elles analysent les conditions environnementales, techniques et humaines qui ont pu conduire à une épidémie, valident ou infirment l'origine hydrique, puis définissent et imposent les mesures de préventions adéquates ;
- 4) d'élaborer, à l'échelle nationale, un historique d'agrégats regroupant l'ensemble de l'information collectée puis de synthétiser l'information en vue d'améliorer les pratiques *via* des dispositions réglementaires ou des actions de promotion des bonnes pratiques.

L'objectif des travaux présentés ici est de (i) poser les fondations d'une méthode simple et rapide de détection d'agrégats de cas de GEAm à l'échelle de la commune ; de (ii) tester cette méthode sur trois départements pilotes ; de (iii) discuter son intérêt et ses limites en vue d'une éventuelle application à la France entière.

## 2 Matériel et méthodes

### 2.1 Données sources

Le nombre de cas de GEAm a été estimé à partir des données de remboursement de médicaments fournies par l'Assurance Maladie (Sniir-AM). Dans une première étape, chaque ordonnance sur laquelle figure au moins un médicament habituellement utilisé dans le traitement des GEA est extraite du Sniir-AM. Ces données sont ensuite traitées automatiquement par un algorithme de discrimination visant à séparer les ordonnances destinées au traitement d'une GEA de celles relevant d'autres pathologies [5]. Cet algorithme se fonde d'une part sur l'occurrence, dans l'ordonnance, de médicaments ou d'associations médicamenteuses spécifiques à certains types de GEA, et d'autre part sur le délai de délivrance des médicaments suite à la prescription. Une enquête auprès des clients de pharmacie [5] a permis d'étalonner les performances de l'algorithme par rapport à la définition clinique de cas de GEA issue d'une conférence de consensus internationale [6]. La sensibilité et la spécificité de l'indicateur de cas de GEAm produit par l'algorithme, par rapport aux cas médicalisés conformes à la définition de consensus, étaient toutes deux estimées à 0,89 en 2010 [5].

### 2.2 Préparation des données

Les cas de GEAm qui résultent du processus de discrimination sont étiquetés au jour de consultation du médecin et à la commune de résidence du bénéficiaire des soins. Les cas dits « touristes », définis par une distance entre la commune de résidence et celle d'exercice du médecin prescripteur supérieure à 50 km, sont exclus.

La période d'étude comprend 208 semaines étendues du 1<sup>er</sup> janvier 2009 au 31 décembre 2012. Les données comprises entre le 21 mars et le 1<sup>er</sup> mai 2011 inclus sont manquantes et les semaines incomplètes en début et en fin de période sont exclues.

Le secteur d'étude inclut 3 départements – le Puy-de-Dôme (63), l'Isère (38) et la Gironde (33) – choisis du fait de la présence d'une équipe locale d'interlocuteurs sensibilisés au sujet. La population totale était de 3,215 millions d'habitants en 2009. Les données de population ont été fournies par l'Institut national de la statistique et des études économiques (Insee). La répartition des communes par taille de population est similaire entre l'Isère et la Gironde (tableau 1). Le Puy-de-Dôme possède moins de communes mais une plus grande proportion de communes de taille inférieure à 500 habitants.

Le nombre de pollutions microbiologiques accidentelles de l'eau distribuée observées sur le secteur d'étude ne diffère pas sensiblement de la moyenne nationale. Le Puy-de-Dôme se situe dans la moyenne, la Gironde est nettement moins touchée et l'Isère nettement plus [7].

**Tableau 1 : Distribution par taille des communes de chaque département du secteur d'étude, exprimée en nombre et en pourcentage, en 2009 (données Insee)**

Taille (habitants)	Gironde		Isère		Puy-de-Dôme		Total	
[0, 100)	14	3 %	23	4 %	29	6 %	66	4 %
[100, 500)	210	39 %	142	27 %	238	51 %	590	38 %
[500, 2 000)	202	37 %	243	46 %	154	33 %	599	39 %
[2 000, 10 000)	94	17 %	109	21 %	39	8 %	242	16 %
[10 000, 50 000)	19	4 %	13	2 %	9	2 %	41	3 %
[50 000 et +	3	1 %	1	0 %	1	0 %	5	0 %
Total	542	100 %	531	100 %	470	100 %	1 543	100 %

Des comptes de cas de GEAm cumulés par semaine ont été utilisés pour plusieurs raisons :

- 1) cela s'accorde avec la durée constatée des épidémies dont l'origine hydrique est reconnue (1 à 3 semaines) ;
- 2) du fait de l'augmentation des effectifs par rapport à des comptages quotidiens, les tests statistiques utilisés sont placés dans de bonnes conditions de validité pour comparer un nombre de cas observé (NCO) à un nombre de cas attendu (NCA) sur des données disponibles à la commune ;
- 3) cela permet de s'affranchir des variations quotidiennes d'incidence dues aux fluctuations d'activité professionnelle, comme les fermetures dominicales, sans qu'il soit nécessaire d'utiliser des variables de contrôle telles que le jour de la semaine ou les jours fériés.

## 2.3 Détection d'agrégats de cas de GEAm

La détection d'agrégats de cas est fondée sur la mise en évidence, *via* l'exploration systématique (*screening*) des communes du secteur d'étude, d'un nombre de cas de GEAm observé (NCO) anormalement élevé par rapport au NCA. Plusieurs méthodes existent à cet effet. Les plus courantes (carte de contrôle, moyennes historiques) ne sont pas adaptées à la nature spatialisée des données et à leur caractère non stationnaire. Le balayage spatiotemporel de Kulldorff [8] répond à l'objectif de détection car il peut prendre en compte la structuration spatiale de réseaux de distribution d'eau destinée à la consommation humaine. Cependant, sa mise en œuvre nécessite des données fiables aujourd'hui indisponibles au niveau national. En effet, les données disponibles pour la France concernant la cartographie des réseaux de distribution ne sont pas suffisamment fiables sur la totalité du territoire pour permettre d'appliquer cette méthode, hormis pour quelques secteurs où elles ont pu faire l'objet de validations spécifiques.

Ainsi, nous avons construit deux méthodes fondées uniquement sur la comparaison entre la situation d'une commune testée et d'un secteur de référence. Elles diffèrent par leur façon d'évaluer le NCA. La mise en œuvre d'une combinaison de ces deux méthodes permet de ne retenir *in fine* que les agrégats de cas de GEAm identifiés par chacune d'entre elles.

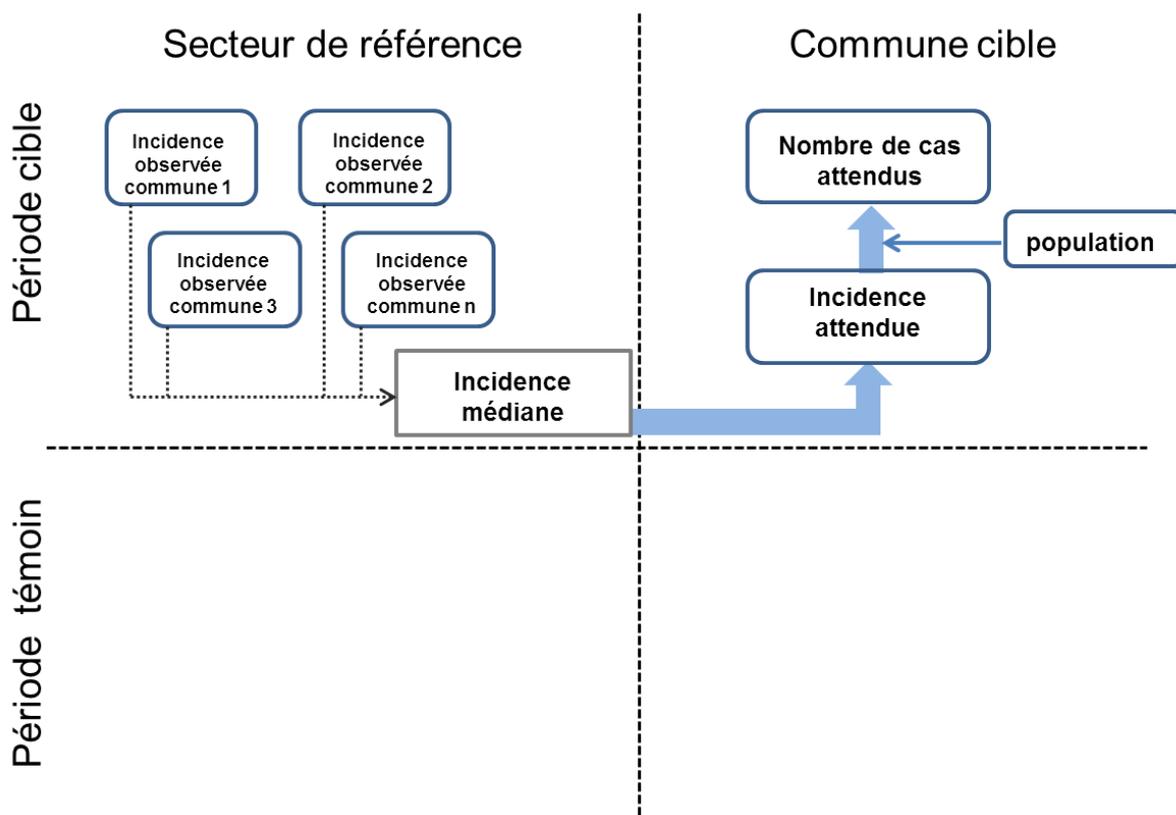
Pour chaque semaine et chaque commune de chaque département, ces méthodes (A et B) sont mises en œuvre, d'abord avec le calcul du NCA puis avec le test de comparaison du NCO et du NCA.

## 2.4 Détermination du nombre de cas attendus

Quelle que soit la méthode, l'intégralité du département d'appartenance de la commune testée est utilisée comme secteur de référence. Ceci permet d'éviter l'utilisation d'une référence géographique différente pour chaque commune testée et de réduire en conséquence le temps de calcul.

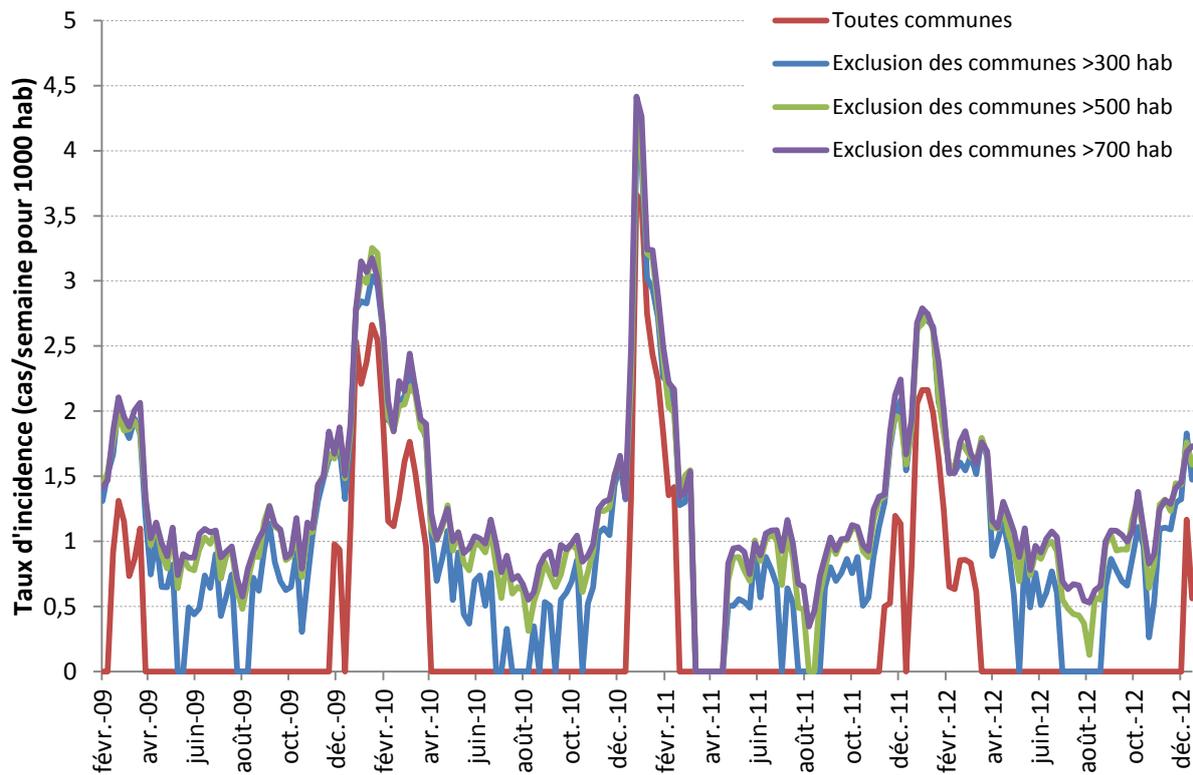
Dans la **méthode A**, pour chaque commune et chaque semaine testée, le NCA est calculé comme le produit du taux d'incidence médian du secteur de référence (TIMref) par la taille de la population de la commune testée (figure 3).

**Figure 3 : Calcul du nombre de cas attendus (NCA) par la méthode A**



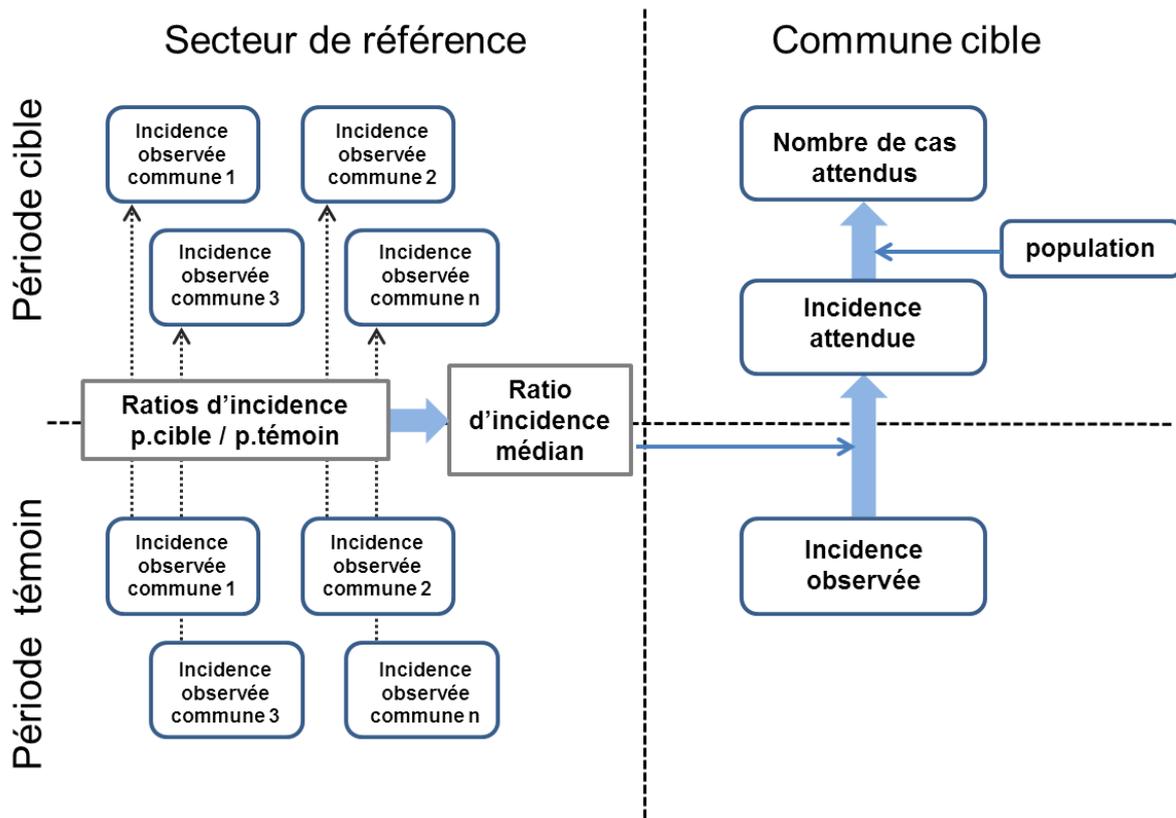
Contrairement au taux d'incidence moyen, le TIMref est moins tiré vers le haut par les fortes incidences des communes en situation épidémique, par les valeurs extrêmes ou aberrantes. Nous supposons donc qu'il est plus à même d'estimer l'incidence hors épidémie d'une zone de référence. Le TIMref peut cependant être nul, empêchant le calcul du NCA pour l'ensemble des communes du secteur de référence, notamment en été et dans les départements ruraux, où la majorité des communes sont de petite taille. Cet effet est évité par l'exclusion des communes de taille inférieure à 500 habitants (figure 4) dans le calcul de TIMref.

**Figure 4 : Effet de l'exclusion des petites communes sur l'incidence communale hebdomadaire médiane dans le Puy-de-Dôme (N communes=470)**



Dans la **méthode B**, un ratio de taux d'incidence est d'abord calculé pour chaque commune entre la semaine cible et la période témoin. Le choix de la période témoin correspond aux quatre semaines comprises entre la cinquième semaine ( $S_{-5}$ ) et la deuxième semaine ( $S_{-2}$ ) précédant la semaine cible ( $S_0$ ). Une semaine tampon est ainsi interposée entre la semaine cible et la période témoin afin d'éviter de baser le calcul du ratio d'incidence sur une semaine intermédiaire durant laquelle aurait pu éventuellement débiter l'épidémie dont on souhaite tester l'existence. Pour chaque commune et semaine testée, l'incidence attendue est calculée comme le produit du ratio d'incidence médian du secteur de référence (RIMref) par l'incidence observée durant la période témoin (figure 5). Comme dans la méthode A et pour des raisons similaires, le RIMref est calculé en excluant les communes de moins de 500 habitants.

Figure 5 : Calcul du nombre de cas attendus (NCA) par la méthode B



Dans les situations où l'incidence observée est nulle durant la période témoin (situation fréquente en été pour les très petites communes), l'incidence attendue sur la semaine testée devient automatiquement nulle, ne permettant pas la réalisation du test de détection d'agrégat. Dans ce cas, le NCO de la période témoin est artificiellement porté à 1 afin de permettre le calcul d'une incidence attendue (par application du RIMref) et l'estimation du NCA sur la période cible. Cette option a tendance à sous-estimer la variation d'incidence réelle entre la période témoin et la période cible (et du NCA) et tend à réduire l'émergence de faux agrégats.

## 2.5 Test du nombre de cas observés contre le nombre de cas attendus

La détection des agrégats hebdomadaires (AH) est fondée sur l'utilisation conjointe de trois critères :

- 1) **Un risque relatif (RR) de GEAm  $\geq 2$**  : estimé par le rapport entre le nombre de cas observés et le nombre de cas attendus. Ce choix est basé sur le constat que des RR de l'ordre de 2 peuvent être observés en dehors de tout facteur de risque d'ordre environnemental entre des populations qui diffèrent uniquement par leur structure en âge ou leur composition socio-économique [9].
- 2) **Un impact sanitaire  $> 5$  cas** : l'impact sanitaire correspond à l'excédent de cas observés par rapport au nombre de cas attendus. Ce choix vise spécifiquement à exclure les agrégats d'origine alimentaire ou par contact qui sont les plus fréquents dans cette catégorie de taille d'agrégats ( $\approx 5$  cas).

- 3) **Une « probabilité » ( $p$ )  $< 1.10^{-5}$**  :  $p$  correspond à la significativité du test statistique réalisé (risque  $\alpha$ ). L'hypothèse nulle est que le NCO sur la commune testée pendant la semaine cible soit égal au NCA. La distribution du nombre de cas de GEAm suit vraisemblablement une loi de Poisson surdispersée. Le test utilisé, pour être formellement interprétable, suppose que cette distribution suive une loi de Poisson. Ainsi, le calcul de la probabilité est exact mais ne permet pas de conclure formellement à la significativité de la différence observée entre le NCO et le NCA. Ce critère est utilisé pour hiérarchiser entre eux les agrégats qui répondent aux critères 1) et 2). Seuls ceux pour lesquels le  $p$  est inférieur à  $1.10^{-5}$  sont retenus pour la suite de l'étude.

## 2.6 Présentation des résultats

Chaque département est d'abord exploré par chaque méthode appliquée séparément. Les AH détectés constituent respectivement les listes A et B. Ensuite, les AH détectés à la fois par la méthode A et la méthode B sont inscrits dans la liste AB (notée  $A \cap B$ ).

Les AH survenus sur une même commune et consécutifs dans le temps sont assemblés en un seul et même agrégat consolidé (AC) dans chacune des 3 listes (A, B et  $A \cap B$ ). Pour caractériser chaque AC, les NCO et NCA sont cumulés sur l'ensemble des AH le constituant. L'impact et le RR sont recalculés sur cette base.

## 2.7 Évaluation de la sensibilité des méthodes par détection d'agrégats simulés

Dans une seconde étape, deux départements (Isère et Puy-de-Dôme) ont été explorés une nouvelle fois par chacune des méthodes après introduction d'agrégats simulés de petite taille dans les séries chronologiques réelles de cas de GEAm. Ces simulations ont été construites à partir de profils d'agrégats d'origine hydrique survenues en France entre 2009 et 2011. Quarante-six agrégats ont ainsi été simulés et greffés sur 24 séries de comptes de cas hebdomadaires communaux dans le département de l'Isère et 22 dans le département du Puy de Dôme.

Deux profils de simulation ont été créés, correspondant à des durées de une et trois semaines, et trois niveaux de taux d'attaques ont été distingués (6, 3 et 1 %). Ces taux correspondant à la fourchette basse des taux d'attaques observés dans les épidémies d'origine hydrique [1] ont été choisis pour tester les capacités des méthodes à détecter des agrégats plutôt petits (tableau 2). Les agrégats simulés ont été introduits sur des communes n'ayant présenté aucun agrégat détecté sur la période 2009-2011.

**Tableau 2 : Distribution des excès de cas et des risques relatifs caractérisant les agrégats simulés afin d'évaluer la sensibilité des méthodes A et B**

	min	percentile			max
		25	50	75	
<b>Nombre de cas de GEAm</b>	6	11	26	58	482
<b>RR</b>	3,2	7,0	9,0	12,5	21,0

Aucun agrégat non détectable par l'une ou l'autre des méthodes (nombre de cas en excès <5 ou RR <2) n'a été simulé. Leur positionnement temporel au sein de la série chronologique a été tiré au sort selon une loi uniforme sur l'ensemble des semaines testables (caractérisées par une absence de données manquantes dans les périodes cibles et témoin).

## 3 Résultats

### 3.1 Fréquence et durée des agrégats

Sur les trois départements d'étude, l'exploration systématique a détecté 826 AH par la méthode A et 543 par la méthode B. 210 AH sont communs aux deux méthodes (liste  $A \cap B$ ), soit 6,5 AH pour 10 000 communes x semaines (CxS) (tableau 3). Le ratio entre le nombre d'AH détectés par A et le nombre d'AH détectés par B est de 1,5 et varie de 2 en Gironde à 0,8 dans le Puy-de-Dôme. La fréquence de détection d'un AH varie entre 5/10 000 CxS en Isère et 8/10 000 CxS dans le Puy-de-Dôme.

**Tableau 3 : Détection d'agrégats de cas de GEAm sur la période 2009-2012 (Puy-de-Dôme, Isère et Gironde) selon la méthode utilisée (A, B ou  $A \cap B$ )**

Secteur (nb communes)	Liste	Agrégats hebdomadaires détectés	Agrégats consolidés détectés	Communes concernées	Nombre de semaines x communes
Gironde (542)	A	348	263	97	112 736
	B	176	165	128	
	$A \cap B$	69	63	54	
Isère (531)	A	339	252	91	110 448
	B	195	190	142	
	$A \cap B$	59	55	48	
Puy-de-Dôme (470)	A	139	126	75	97 760
	B	172	158	105	
	$A \cap B$	82	75	60	
Total (1 543)	A	826	641	263	320 944
	B	543	513	375	
	$A \cap B$	210	193	162	

*Agrégats consolidés = agrégats hebdomadaires agglomérés si consécutifs dans le temps.*

La règle de l'intersection ( $A \cap B$ ) opère une sélection drastique puisque seuls 18 % des AH détectés par l'une ou l'autre méthode ( $A \cup B$ ) se retrouvent *in fine* dans la liste  $A \cap B$ . Son effet est variable selon le département : dans le Puy-de-Dôme, 36 % des AH détectés par l'une des 2 méthodes se retrouvent dans l'intersection contre 12 % dans l'Isère.

Les 210 AH détectés sur l'ensemble du secteur d'étude correspondent à 193 AC présents sur la liste  $A \cap B$ . Seuls 20 % des AC détectés par l'une ou l'autre des méthodes A ou B se retrouvent *in fine* dans la liste  $A \cap B$  (tableau 3). La méthode A détecte 1,3 fois plus d'AC que la méthode B sur le secteur d'étude. Le rapport du nombre d'AC de la liste A sur le nombre d'AC de la liste B varie de 1,6 en Gironde à 0,8 dans le Puy-de-Dôme.

Seuls 16 % des AC détectés par  $A \cap B$  se sont produits sur des petites communes (moins de 500 hab.) et moins de 2 % sur de très grandes communes de 50 000 hab. et plus (tableau 4). Un seul agrégat a été

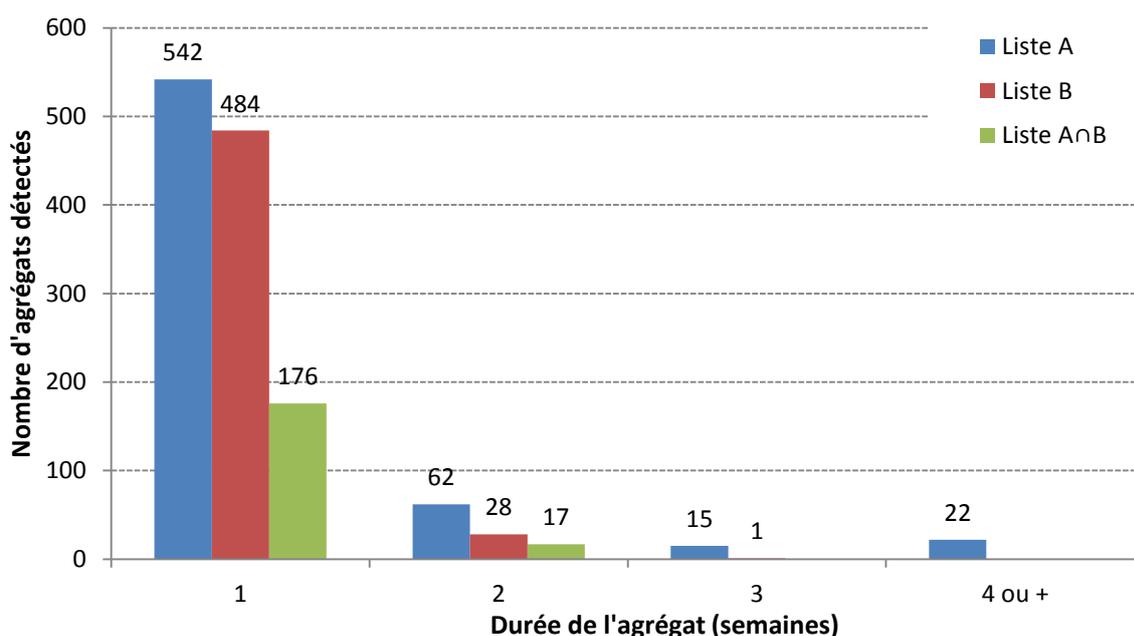
détecté par la méthode A sur une très petite commune (moins de 100 hab.) et aucun par la méthode B. La répartition du nombre d'AC détectés est globalement similaire entre les méthodes A et B excepté sur la tranche 10 000-50 000 hab. sur laquelle la méthode A détecte plus d'AC. La probabilité de détection d'un AC par la méthode  $A \cap B$  est nulle sur les très petites communes et augmente en fonction de la taille de la commune analysée (tableau 4), excepté pour les très grandes communes.

**Tableau 4 : Distribution des agrégats consolidés détectés par strate de population sur la période 2009-2012 (Puy-de-Dôme, Isère et Gironde)**

Taille (habitants)	Nombre total de communes	Liste A		Liste B		Liste $A \cap B$		Communes positives pour au moins un agrégat ( $A \cap B$ )	
		Nombre Agrégats	%	Nombre Agrégats	%	Nombre Agrégats	%	Nombre	%
[0,100)	66	1	0 %	0	0 %	0	0 %	0	0,0
[100,500)	590	81	13 %	47	9 %	30	16 %	27	4,6
[500,2 000)	599	192	30 %	238	46 %	70	36 %	58	9,7
[2 000,10 000)	242	171	27 %	182	35 %	67	35 %	60	24,8
[10 000, 50 000)	41	192	30 %	41	8 %	23	12 %	16	39,0
[50 000 et +	5	4	1 %	5	1 %	3	2 %	1	20,0
Total	1 543	641	100 %	513	100 %	193	100 %	162	10,5

Plus de 9 AC sur 10 (91 %) de la liste  $A \cap B$  durent une seule semaine. Le reste des AC (17/193) durent 2 semaines. Aucun agrégat de la liste  $A \cap B$  n'avait une durée supérieure à 2 semaines. Ces proportions étaient proches pour la méthode B (un seul agrégat de 3 semaines) alors que 37 agrégats (6 %) détectés par la méthode A sont d'une durée de 3 semaines ou plus (figure 6), avec un maximum de 11 semaines (Isère en juin 2012).

**Figure 6 : Distribution des agrégats consolidés de cas de GEAm selon leur durée (Puy-de-Dôme, Isère et Gironde ; 2009-2012)**



## 3.2 Communes à risque chronique

Sur les 4 années de la période d'étude, 10,5 % des communes du secteur d'étude ont été confrontées au moins une fois à un agrégat de cas de GEAm détecté par les méthodes A et B (liste  $A \cap B$ ) (tableau 4). La méthode B détecte des AC sur un nombre plus élevé de communes que la méthode A, quel que soit le département (tableau 3).

Les 193 AC identifiés sur la liste  $A \cap B$  sur l'ensemble du secteur d'étude concernent 162 communes différentes. La grande majorité de ces communes (86 %) est concernée par un seul AC sur la période d'étude. Moins de 1 commune sur 7 positives (avec au moins 1 agrégat) ont connu l'occurrence de 2 AC ou plus (tableau 5). Un maximum de répétitions de 4 AC par commune a été observé sur la liste  $A \cap B$ , à comparer avec un maximum de 32 AC par commune selon la seule méthode A et 5 AC par commune selon la seule méthode B.

**Tableau 5 : Répartition des communes positives selon le nombre d'agrégats consolidés détectés (Puy-de-Dôme, Isère et Gironde, N communes=1 543) ; 2009-2012**

Nb d'agrégats consolidés par commune	Liste A		Liste B		Liste $A \cap B$	
	Nombre de communes	%	Nombre de communes	%	Nombre de communes	%
1	164	62 %	270	72 %	139	86 %
2	47	18 %	79	21 %	16	10 %
3-5	30	11 %	26	7 %	7	4 %
6 ou +	22	8 %	0	0 %	0	0 %
Total	263	100 %	375	100 %	162	100 %

Les communes concernées par la présence d'AC détectés par A et B sont pour moitié isolées géographiquement ou bien tendent à former des petits groupes de 2 à 8 communes. La répartition géographique des communes positives ou des groupes de communes positives ne semble pas structurée en Gironde et en Isère. Elle se concentre au nord du département dans le Puy-de-Dôme (annexes A1, A2 et A3).

## 3.3 Nombre de cas impliqués

Deux AC sur 5 détectés sur la liste  $A \cap B$  impliquent moins de 10 cas de GEAm. Ces petits agrégats concentrent 16 % de l'ensemble des cas détectés (tableau 6). À l'opposé, 1 AC sur 15 impliquent 50 cas et plus et représentent 28 % des cas totaux de GEAm. Le plus grand AC détecté implique 199 cas sur une durée de 2 semaines.

**Tableau 6 : Distribution des agrégats consolidés et détectés par la méthode A∩B selon le nombre de cas de GEAm impliqués et caractérisation de l'impact de ces mêmes agrégats par la seule méthode A (Puy-de-Dôme, Isère et Gironde ; 2009-2012)**

Impact sanitaire (nombre de cas)	Nombre d'agrégats consolidés	%	Total cas	%	Total cas estimés par la méthode A	%
[6,10)	79	41 %	533	16 %	545	13 %
[10,20)	68	35 %	932	28 %	872	20 %
[20,50)	35	18 %	927	28 %	1 165	27 %
[50,100)	9	5 %	584	17 %	495	12 %
[100 ou +	2	1 %	356	11 %	1 194	28 %
Total	193	100 %	3 332	100 %	4 271	100 %

La méthode B montre une tendance à sous-évaluer la durée des AC détectés (figure 6) par rapport à la méthode A et donc également leur impact sanitaire. Ainsi, 58 % des agrégats détectés par la méthode B présentent moins de 10 cas de GEAm contre seulement 29 % des agrégats détectés par A (tableaux 6a et 6b). Cette sous-évaluation des impacts par la méthode B se retrouve dans les estimations d'impact des AC présents sur la liste A∩B. Lorsque ces mêmes agrégats sont caractérisés par la seule méthode A, on constate que l'estimation du nombre de cas impliqués dans les agrégats de 100 cas ou plus est 3 fois supérieure (tableau 6a et 6b).

**Tableau 6a : Distribution des agrégats consolidés et détectés par la méthode A selon le nombre de cas de GEAm impliqués (Puy-de-Dôme, Isère et Gironde ; 2009-2012)**

Impact sanitaire (nb de cas)	Nombre agrégats consolidés	%	Total cas	%
[6,10)	186	29 %	1 342	8 %
[10,20)	212	33 %	3 004	18 %
[20,50)	179	28 %	5 133	31 %
[50,100)	44	7 %	2 918	18 %
[100 ou +	20	3 %	4 253	26 %
Total	641	100 %	16 650	100 %

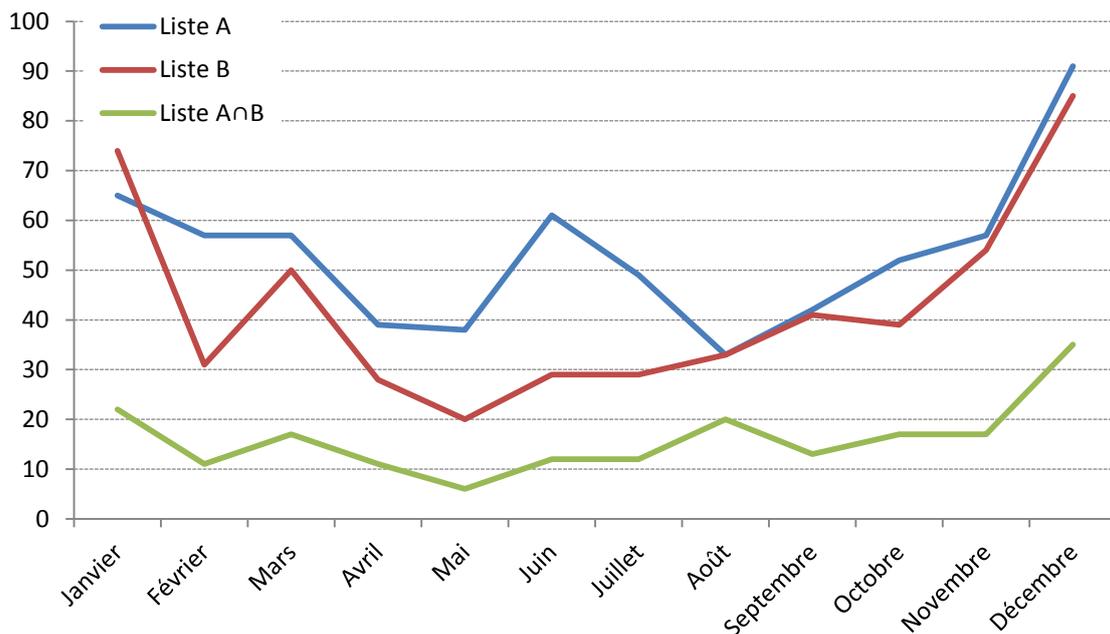
**Tableau 6b : Distribution des agrégats consolidés et détectés par la méthode B selon le nombre de cas de GEAm impliqués (Puy-de-Dôme, Isère et Gironde ; 2009-2012)**

Impact sanitaire (nb de cas)	Nombre agrégats consolidés	%	Total cas	%
[6,10)	299	58 %	2 084	31 %
[10,20)	146	28 %	2 008	30 %
[20,50)	54	11 %	1 436	21 %
[50,100)	10	2 %	676	10 %
[100 ou +	4	1 %	598	9 %
Total	513	100 %	6 802	100 %

### 3.4 Variations saisonnières de la détection

Le nombre d'AC détectés sur la liste  $A \cap B$  est de 16 par mois en moyenne ; il évolue entre 6 en mai et 35 en décembre (figure 7). Quelle que soit la méthode utilisée, deux fois plus d'agrégats sont détectés durant le mois de décembre, soit du fait d'un artéfact lié aux méthodes, soit du fait de conditions naturelles (cf. Discussion). Les variations saisonnières sont similaires entre le nombre d'agrégats présents sur la liste  $A \cap B$  et ceux détectés par chacune des deux méthodes A ou B.

**Figure 7 : Distribution par mois du nombre total d'agrégats consolidés détectés sur la période 2009 à 2012 (Puy-de-Dôme, Isère et Gironde)**



### 3.5 Évaluation de la sensibilité des méthodes

Les épidémies de GEA d'origine hydrique recensées par l'InVS sur la base des signalements par les acteurs de terrain sont caractérisées par un nombre de cas de GEA important (plus de 100 cas de GEA, correspondant à plus de 30 cas de GEAm). L'aptitude des méthodes A et B à identifier ces épidémies a été vérifiée. Les 4 épidémies recensées entre 2009 et 2012 sur le secteur d'étude ont été identifiées.

Les simulations d'agrégats donnent une indication supplémentaire sur la sensibilité de la méthode  $A \cap B$ . Les contraintes imposées sur l'impact minimum (>5 cas de GEAm surnuméraires), conduisent toutefois à une sous-représentation des petites communes parmi les simulations : 9 agrégats simulés concernent des communes entre 100 et 500 habitants, 36 des communes entre 500 et 10 000 habitants et une seule simulation concerne une commune urbaine. Sur ces 46 AC simulés, 37 (80 %) sont identifiées sur la liste  $A \cap B$ . Les 9 faux négatifs correspondent à des agrégats de petite taille (6 sur 9 sont de taille inférieure à 10 cas contre 29 % des AC totaux) et majoritairement hivernaux (6 sur 9). La méthode B est à la source du

manque de sensibilité, puisqu'elle ne détecte pas 8 des AC non détectés par  $A \cap B$ , alors que la méthode A seule repère 5 de ces 9 AC. Trois AC ne sont détectés ni par A ni par B. Les 9 faux négatifs sont petits (max = 21 cas de GEAm et p75 = 10 cas de GEAm) et majoritairement hivernaux (6/9 en janvier ou février).

Les 46 AC simulés correspondent à 80 AH simulés, dont seulement 58 (72,5 %) sont détectés sur la liste  $A \cap B$  (tableau 7). Le protocole utilisé fixe la spécificité à 1 du fait de l'injection d'agrégats dans des communes pour lesquelles aucun agrégat n'avait été précédemment détecté.

**Tableau 7 : Performances de détection des agrégats hebdomadaires simulés par les méthodes A et B (liste  $A \cap B$ ) pour 46 communes de l'Isère et du Puy-de-Dôme entre 2009 et 2012**

		Nombre d'agrégats hebdomadaire		Total
		détectés	non détectés	
Nombre de semaines	avec agrégat simulé	58	22	80
	sans agrégat simulé	0	9 488	9 488
Total		58	9 510	9 568

### 3.6 Impact national attendu

Dans la mesure où la fréquence des pollutions microbiologiques accidentelles de l'eau distribuée observée dans les départements d'étude ne diffère pas de la moyenne nationale [7], nous avons supposé que la fréquence des agrégats de GEAm observée dans l'étude était représentative du niveau national et inféré la fréquence observée dans l'étude au niveau national. 1 600 AC impliquant 5 cas de GEAm ou plus pourraient ainsi être détectés annuellement en France et 950, 380 et 90 AC impliquant respectivement 10, 20 et 50 cas de GEAm ou plus. Ces ordres de grandeurs seront à valider et à préciser dans le futur.

## 4 Discussion

### 4.1 Comparaison des deux méthodes de détection A et B

Les deux méthodes développées (A et B) présentent des complémentarités de détection qui permettent, *via* une méthode assemblée ( $A \cap B$ ), de palier les défauts de chacune des méthodes prises isolément (tableau 8).

**Tableau 8 : Caractéristiques comparées des méthodes A et B de détection d'agrégats de GEAm**

Situation	Méthode A	Méthode B
Données manquantes	Calcul impossible Effet limité aux semaines cibles en « données manquantes »	Calcul impossible Effet élargi aux périodes témoins en « données manquantes »
Erreur sur le code Insee des données sources	Grande sensibilité aux faux positifs avec répétitions possibles	Faible sensibilité car comparaison entre période témoin et période cible
Variations saisonnières d'incidence des GEA	Effet +/- contrôlé. Faux positifs possibles en début d'épidémie hivernale (décembre), par exemple, si l'épidémie hivernale commence sur la commune cible, avant diffusion à l'ensemble du département	
Différences locales de l'incidence des GEAm sans lien avec l'exposition (accès au soin, critères socio-économiques)	Effet non contrôlé Possibilité de faux positifs et de répétitions	Effet contrôlé car comparaison entre période témoin et période cible
Épidémie de longue durée ( $\geq 3$ semaines)	Effet contrôlé	Risque de faux négatif croissant avec la durée de l'épidémie ou risque de sous-estimation de l'impact d'une épidémie détectée. Effet contrôlable <i>a posteriori</i>
Répétition d'agrégats sur une même commune	Effet contrôlé	Effet non contrôlé : risque de faux négatif croissant avec le nombre de répétitions
Taux d'incidence communal majoritairement nuls en secteur témoin, semaine cible (pour les très petites communes en été)	Risque de faux positifs. Mesure palliative : exclusion des petites communes pour le calcul du TIMref	
NCO nul sur la période témoin pour la commune cible	Sans objet	Risque de faux positifs. Mesure palliative : imputation du NCO à 1

La méthode B apparaît plus sensible à la présence de données manquantes du fait de l'intervention d'une période témoin dans le calcul du nombre de cas attendus. Ainsi, les 5 premières semaines de la période d'étude et les 5 semaines consécutives aux semaines de données manquantes d'avril 2011 ne peuvent pas être testées avec la méthode B, alors qu'elles peuvent l'être avec la méthode A.

L'utilisation d'une période témoin pour déterminer le nombre de cas attendus hors épidémie peut être problématique lorsqu'un agrégat de cas se produit durant cette période témoin. Ce biais n'affecte pas la capacité de la méthode B à détecter le début d'une épidémie sporadique mais conduit à la sous-estimation

de la durée des agrégats et donc de leur impact global. Au-delà de la deuxième semaine de l'agrégat, la période témoin inclut mécaniquement le début de l'épidémie. La détection d'agrégats répétitifs à moins de 5 semaines d'intervalles sur une même commune peut toutefois être mise à mal car la période témoin utilisée pour détecter le second agrégat inclus mécaniquement le premier, ce qui a pour conséquence de diminuer la probabilité de détection du deuxième agrégat. Ce biais est spécifique à la méthode B qui est la seule à utiliser une référence temporelle. Pour pallier à ce défaut, il est envisageable de caractériser l'impact des agrégats détectés sur la liste  $A \cap B$  *via* la méthode A seule (tableau 6).

L'utilisation d'une référence purement géographique par la méthode A conduit à la sur-détection d'agrégats sur les communes de plus de 10 000 habitants (tableau 4), défaut compensé par l'association avec la méthode B, insensible à ce biais. Les communes de plus de 10 000 habitants, du fait d'une population plus jeune, peuvent présenter un nombre de cas de GEAm qui peut apparaître anormalement élevé lorsqu'il est comparé à une référence géographique dont la structure en âge diffère (en l'occurrence plus vieille). La standardisation sur l'âge des cas de GEAm est une solution qui permettra d'éviter cette sur-détection.

De plus, les possibles erreurs d'attribution géographique aux cas de GEAm, par exemple par substitution d'un code Insee par un code postal, conduisent à une surincidence virtuelle de cas de GEAm sur les communes sur lesquelles les cas sont faussement attribués et entraîne un excès artificiel d'agrégats. Ces erreurs, probablement conséquentes à une erreur de saisie des caractéristiques propres à certains bénéficiaires des soins, ne peuvent être corrigées qu'en amont de l'alimentation des bases de données du Sniir-AM. Elles restent toutefois assez rares.

Concernant la saisonnalité des agrégats détectés, deux types de raisons peuvent expliquer la concentration des agrégats en décembre : les conditions naturelles favorables ou un artéfact méthodologique. Ces deux raisons sont compatibles.

Les conditions naturelles favorables couvrent la reprise des ruissellements hivernaux sur les terres agricoles et la remise en circulation dans l'environnement hydrique des virus fécaux dans le contexte de l'épidémie hivernale.

L'épidémie hivernale d'étiologie virale qui touche l'hémisphère nord entre décembre et février se propage principalement par contact interpersonnel ou *via* les objets contaminés. Lors de l'arrivée de l'épidémie, la dynamique de l'incidence est moins lissée au niveau local (figure 2) qu'aux niveaux national ou départemental. Il est possible que la survenue brutale de cas sur une commune soit la conséquence de l'arrivée de l'épidémie hivernale et non d'une contamination microbiologique de l'eau distribuée. Par construction, les méthodes A et B permettent un certain contrôle de l'effet saisonnier. Cependant, ce contrôle est imparfait car il repose sur une comparaison géographique hebdomadaire au niveau départemental et peut ne pas couvrir les effets très localisés de la montée en puissance de l'épidémie, qui surviennent principalement au mois de décembre.

## 4.2 Imputation des agrégats à l'eau : une question ouverte

L'évaluation de la sensibilité et de la spécificité de la méthode de détection d'agrégats ( $A \cap B$ ) est subordonnée à une séparation claire des agrégats visés par la méthode (événements ayant pour origine la contamination microbienne de l'eau du robinet) des autres événements dont la détection n'est pas souhaitée, à savoir :

- les agrégats liés au hasard, c'est-à-dire sans cause clairement identifiée ;
- les faux agrégats liés à des biais méthodologiques ;
- les agrégats véhiculés par contact et s'inscrivant dans l'épidémie virale hivernale ;
- les toxi-infections alimentaires collectives ;
- les baignades en piscine ou en milieu naturel.

Les contraintes liées au travail de terrain au sein des ARS plaident en faveur de la mise au point d'une méthode de détection qui privilégie la spécificité à la sensibilité. Au niveau local, l'enquête de terrain devrait permettre d'imputer ou non un agrégat détecté à l'eau du robinet et d'identifier les actions de prévention adéquates. Cette enquête représente une mobilisation de moyens non négligeable. Si l'occurrence de « faux négatifs » n'a pas d'enjeu en termes de travail de terrain, l'investigation de « faux positifs » rendra vains les efforts engagés et discréditera le système de détection auprès des acteurs de terrain.

Dans cet esprit, les agrégats rapportés aux acteurs de terrain seront classés selon le nombre de cas impliqués (impact sanitaire) et de répétitions afin d'encourager l'affectation des moyens locaux à l'investigation des agrégats correspondant aux impacts majeurs en termes de santé publique. *A contrario*, les événements isolés et de faible impact (par exemple, moins de 10 cas de GEAm) pourraient, dans un premier temps, être considérés comme moins prioritaires pour la mise en œuvre d'investigations par les acteurs de terrain. La recherche d'une meilleure sensibilité de la détection n'est donc pas une priorité à court terme.

Les simulations d'épidémies réalisées ont toutefois visé à estimer la sensibilité des méthodes de détection utilisées. À partir des courbes épidémiques à l'échelle de la commune et de la semaine, il est difficile de distinguer clairement une épidémie d'origine hydrique d'une autre épidémie. Dès lors, il s'agit de faire en sorte que la méthode de détection favorise *a priori* l'émergence d'agrégats d'origine hydrique plutôt que d'une autre origine, et de fournir à l'investigateur des indicateurs qui aident à leur discrimination.

Les critères de discrimination des agrégats constituent une première barrière contre la détection de faux positifs. Si on assume la distribution de Poisson des comptes de cas hebdomadaires communaux, la multiplication des tests statistiques opérés engendre sur chacun des départements du secteur d'étude un faux positif tous les 4 ans. La condition de sélection  $RR > 2$  devrait par ailleurs prémunir la méthode A contre la survenue de faux positifs attribuables aux variations intercommunales du taux d'incidence des GEA induites par des facteurs socio-économiques, comme l'accès au médecin ou le niveau d'éducation [10].

La combinaison de la méthode A avec la méthode B, qui est insensible à ce biais, est une barrière supplémentaire contre la détection de ces faux positifs dans la liste  $A \cap B$ . Nous observons plus généralement une complémentarité des méthodes A et B contre la survenue de faux positifs. La faille qui subsiste dans la lutte contre les faux positifs touche à la possibilité de confusion, en décembre, entre une poussée communale de l'épidémie hivernale de gastro-entérites et un agrégat d'origine hydrique.

Le risque de confusion avec une toxi-infection alimentaire collective (Tiac) excluant l'eau ne peut pas non plus être éliminé, mais la conséquence d'une méprise est moins grave dans le sens où la découverte *a posteriori* d'une Tiac peut présenter en soi un intérêt de santé publique.

Un peu plus de 1 200 Tiac sont rapportées chaque année en France [11]. Les Tiac impliquent un nombre de cas de GEAm généralement réduit mais qui varie selon que la restauration soit de type « familiale » ou de type « collective » (tableau 9). Les foyers familiaux (32 %) impliquent trop peu de cas pour être investigués (moins de 1 % des Tiac atteignent 25 cas de GEA ou plus soit environ 8 cas de GEAm ou plus). Par ailleurs, la majorité des établissements de restauration collective s'adressent à des classes d'âge particulières : essentiellement des enfants pour la restauration scolaire ou périscolaire, des adultes pour la restauration d'entreprise, des personnes âgées pour les maisons de retraite. Les Tiac survenant dans ces établissements sont donc *a priori* identifiables par la structure d'âge des cas (par exemple par le ratio du nombre de cas pédiatriques sur le nombre de cas adultes).

**Tableau 9 : Distribution des Tiac en fonction du nombre de cas et par lieu de contamination probable, et possibilité de discrimination des Tiac par la distribution par âge des cas (par rapport à des agrégats d'origine hydrique), France, 2009-2012 (source : InVS)**

Nombre de malades	Restauration familiale	Restauration commerciale	Restauration scolaire	Inst. Med. Soc.*	Restauration d'entreprise	Autres collectivités**	Foyers diffus***	Lieu non précisé	Total
2-9	1 405	1 484	164	137	119	218	14	26	3 567
10-24	80	93	171	147	38	168	9	7	713
25-49	13	24	77	64	24	88	4	2	296
50 à 99	1	8	21	11	10	28	1	1	81
≥100	-	2	14	2	2	11	-	-	31
Total	1 499	1 611	447	361	193	513	28	36	4 688
Discrimination par l'âge	Non	Non	Oui	Partielle	Oui	Partielle	Sans objet	Partielle	

\*Institutions médico-sociales : hôpitaux, établissements d'accueil de personnes âgées, crèches.

\*\*Banquets, kermesses, prisons.

\*\*\*Tiac au cours de l'investigation desquelles des foyers ont été déclarés dans plusieurs endroits simultanément et dont l'aliment responsable provient du même lot.

Si on ne retient que les Tiac de plus de 25 cas (soit en moyenne 8 cas de GEAm) survenant en population générale ou en établissement dont la fréquentation concerne toutes les classes d'âge (par exemple les hôpitaux), seuls une quarantaine de foyers rapportés chaque année en France (soit 3 % des Tiac) pourraient être confondus avec des agrégats d'origine hydrique.

Parmi les informations disponibles dans le Sniir-AM, la prescription de coproculture pourrait être utilisée pour aider au repérage *a priori* des Tiac. Sept pourcents des cas de GEAm se voient prescrire une

coproculture [12]. Cette prescription évoque une suspicion d'origine bactérienne de la part du médecin, comme le montre la similitude des séries chronologiques de prescription des coprocultures et celles des déclarations obligatoires de salmonelloses (Van Cauteren, communication personnelle). Or, environ 80 % des Tiac (foyers confirmés ou suspectés) sont d'origine bactérienne, tandis que les épidémies d'origine hydrique ont une étiologie bactérienne dans seulement 22 % (France [1]) à 33 % (Angleterre et Pays de Galles [13]) des cas. L'utilité des données de coproculture pour discriminer les Tiac des épidémies d'origine hydrique reste cependant à confirmer car les prescriptions peuvent varier proportionnellement dans le temps (moins en hiver et plus en été).

Du fait du système de déclaration obligatoire (DO), une partie des Tiac auront déjà été identifiée quand les données d'agrégats issues de la détection automatisée seront transmises à l'autorité sanitaire locale. Ces Tiac ne seront donc pas à la source d'inutiles investigations sur l'origine hydrique. En revanche, le risque existe pour les Tiac non rapportées. L'exhaustivité de la DO des Tiac à salmonelles a été estimée en 2000 à 26 % [IC<sub>95%</sub>: 22-31] [14], ce qui laisse augurer d'un grand nombre de Tiac non découvertes. Le taux de Tiac non détectées devrait toutefois sensiblement diminuer quand le nombre de cas impliqué augmente, du fait d'une meilleure détectabilité. On peut globalement attendre des fréquences d'occurrence des Tiac et des épidémies d'origine hydrique du même ordre de grandeur parmi les agrégats présentés aux autorités locales.

Deux types d'arguments peuvent enfin être invoqués pour attribuer un agrégat à l'eau distribuée : l'emprise spatiale de l'agrégat et l'identification de conditions défavorables d'exploitation de l'eau.

Pour les agrégats concomitants affectant plusieurs communes, l'adéquation spatiale entre le secteur épidémique et l'unité de distribution (UD) est un argument très fort en faveur de l'origine hydrique. Cette configuration ne concerne cependant que 15,5 % des UD (tableau 10). *A contrario*, quand l'emprise de l'UD correspond à une commune ou à une partie d'une commune, l'argument de la spatialité de l'agrégat est sans objet.

**Tableau 10 : Distribution du nombre de communes par unités de distribution (UD), Sise-eaux, 2014**

Communes par UD	Gironde		Isère		Puy-de-Dôme		Secteur d'étude	
1	61	46,9 %	778	90,2 %	480	84,7 %	1 319	84,6 %
2	15	11,5 %	36	4,2 %	41	7,2 %	92	5,9 %
3 ou +	54	41,5 %	49	5,7 %	46	8,1 %	149	9,6 %
Total	130	100 %	863	100 %	567	100 %	1 560	100 %

L'enquête environnementale vise à mettre en évidence, dans la période pré-épidémique, des événements susceptibles de plaider en faveur d'une origine hydrique de l'épidémie [15] :

1. conditions météo-hydrologiques défavorables (pluies intenses, crue, pollution de la ressource (turbidité), débordements de systèmes d'assainissement en amont du captage, etc.) ;
2. dysfonctionnements dans l'exploitation (panne du système de désinfection, clarification défectueuse, travaux sur réseau...) ;

3. plaintes d'usagers (odeur, goût, couleur de l'eau) ;
4. résultat non conforme du contrôle sanitaire de l'eau.

La concomitance de l'un de ces événements et d'un agrégat de cas de GEAm constitue une quasi-preuve de l'origine hydrique de l'agrégat.

### 4.3 Information utile aux investigateurs de terrain

Les données qui seront proposées aux autorités locales de santé sont la liste des agrégats consolidés présent sur la liste  $A \cap B$  et la liste des communes impactées. Ces listes sont ordonnées par impact décroissant, c'est-à-dire priorisées en termes de santé publique, en vue de diriger les investigations de terrain sur les événements les plus importants et, plus encore, sur les communes à agrégats répétitifs.

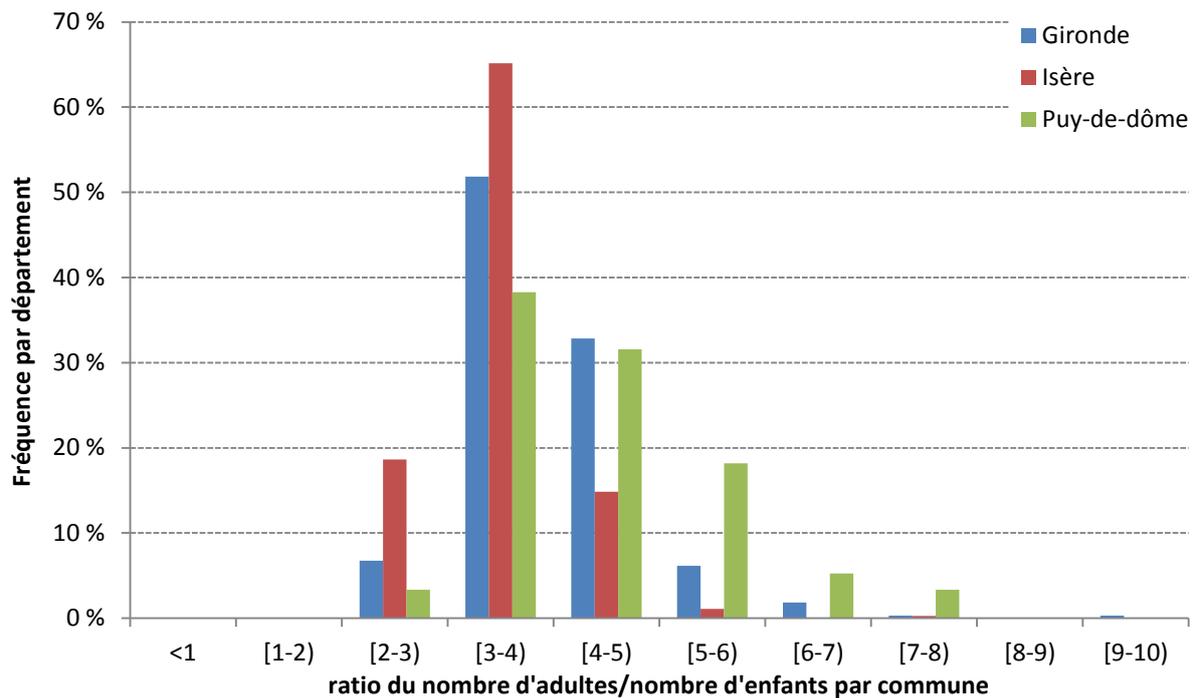
Les « probabilités »  $p$  qui entrent dans le procédé des méthodes de détection ne sont pas exactes. Elles possèdent une valeur relative qui permet de les utiliser comme critère d'identification d'agrégats mais il n'y a pas d'intérêt à les présenter aux autorités locales. L'estimation des RR est très sensible à la méthode ( $A$  vs.  $B$ , par exemple, résultats non montrés). Il est préférable de leur substituer une estimation simple, vérifiable par l'utilisateur, et indépendante de la méthode, comme le ratio entre les taux d'incidence du secteur cible et du secteur de référence.

Nous avons identifié un risque de sous-estimation par la méthode  $B$  (et donc par  $A \cap B$ ) de la durée des épidémies et par conséquent de leur impact. L'estimation produite par la méthode  $A$  peut être proposée en remplacement afin de pallier ce problème. L'expérience montre que ce choix triple le nombre de cas attachés aux agrégats consolidés de plus de 100 cas, avec une conséquence potentiellement importante sur la probabilité d'investigation des événements sous-jacents.

Augmenter le pouvoir de discrimination du système vis-à-vis des Tiac, et des agrégats d'origine non hydrique en général, constitue un enjeu important et peut être conforté par l'ajout d'indicateurs dédiés.

Le rapport des effectifs entre les cas enfants et adultes (1-15 ans vs.  $\geq 16$  ans) peut mettre en évidence une Tiac en communautés d'enfants ou de personnes âgées et doit apparaître parmi les données fournies aux partenaires de terrain. La distinction des effectifs enfants et adultes permettra aussi de standardiser *a minima* le calcul du risque sur la structure en âge de la population. Au niveau communal, le ratio du nombre d'adultes sur le nombre d'enfants varie de 2 à 10 dans les communes de 500 habitants ou plus (figure 8). Dans les communes de moins de 500 habitants, ce ratio est généralement plus élevé (jusqu'à 34 dans le Puy-de-Dôme). La standardisation permettra également de tenir compte d'un risque de GEAm près de 10 fois plus élevé chez les enfants que chez les adultes [3].

**Figure 8 : Distribution des ratios Adultes/Enfants de chaque commune (de 500 habitants ou plus) pour les 3 départements du secteur d'étude (données Insee 2010)**



La présence de cas ayant reçu une prescription de coproculture par le médecin constitue aussi une indication intéressante pour l'investigateur (fréquence au sein de l'agrégat et test de l'hypothèse d'une fréquence similaire à la moyenne nationale).

La méthode décrite ne tient compte que des cas présents sur leur lieu de résidence au moment de l'évènement. En revanche, la démographie touristique n'est pas connue à l'échelle CxS. En faisant l'hypothèse qu'en ce qui concerne les vacances scolaires, elle est plus ou moins conservée d'une année sur l'autre, on peut alors mettre en œuvre la méthode des moyennes historiques qui procède de la comparaison directe d'effectifs de malades et non de taux d'incidence. Le nombre de cas de GEAm observés durant une semaine de vacances scolaires peut être testé contre celui observé durant les autres années pendant la semaine de vacances scolaires la plus similaire. Cette méthode reste approximative mais elle constitue le seul recours pour traiter les épidémies qui frappent les touristes, et qui pourraient avoir un impact important en France [16].

L'information de terrain est prépondérante dans l'estimation du degré de vraisemblance de l'origine hydrique des agrégats détectés. Automatiser le processus d'imputation à l'eau représenterait un effort important pour une conclusion qui resterait incertaine pour de nombreuses épidémies et devrait systématiquement être validée sur le terrain. Cela nécessiterait notamment la connaissance des limites d'UD et leur mise à jour régulière. Cette information n'est, à ce jour, pas connue de façon fiable au niveau national. De plus elle ne permettrait pas de trancher dans la totalité des situations, du fait de l'existence d'interconnexions fréquemment modifiées sur le terrain. Enfin, le point d'introduction de la pollution se situe

dans le réseau de distribution dans la moitié des épidémies d'origine hydrique étudiées [15]. La contamination peut alors ne pas toucher toute l'unité de distribution.

L'imputation à l'eau nécessite un changement de paradigme par rapport à l'approche classique, qui repose sur le signalement aux autorités locales et des investigations épidémiologiques rapides, incluant notamment des prélèvements d'eau en urgence et des analyses de selles chez les malades [15]. Dans la situation actuelle, la preuve de l'origine hydrique est réputée suffisante si on retrouve la souche pathogène dans l'eau et dans les selles des malades ou si une étude épidémiologique analytique montre un risque croissant proportionnellement à la consommation d'eau du robinet [17]. Une surveillance du risque épidémique d'origine hydrique fondée sur la détection automatisée et la réalisation d'enquêtes environnementales rétrospectives ne satisfait pas à ces critères, qui devront être revus à la lumière de l'expérience.

## 5 Conclusion

À l'issue de la phase pilote, il apparaît que la recherche automatisée d'agrégats de cas de GEAm est faisable et pourrait mettre en évidence annuellement plusieurs centaines d'agrégats de 5 cas de GEAm ou plus (soit 15 cas de GEA au minimum) en France chaque année, contre deux épidémies d'environ 100 cas de GEA actuellement sur la base de la surveillance actuelle [1].

Dans la méthode présentée, la sélection des agrégats de cas de GEAm repose sur leur détection par deux tests complémentaires. Cette démarche offre un bon rempart contre l'occurrence de faux positifs. Les défauts de spécificité résiduels concernent le début de la période épidémique hivernale, avec un risque de faux positif, et la possibilité de confusion, permanente mais modérée, avec une Tiac de grande ampleur non repérée par ailleurs. Ce dernier aspect, vu comme un défaut dans le cadre de l'objectif poursuivi, présente cependant en soi un intérêt en termes de santé publique.

Préalablement à l'installation d'un processus national de détection automatisé d'agrégats de cas de GEAm, le choix définitif de la méthode à mettre en œuvre doit se baser sur un ensemble de critères rigoureux propres à la méthodologie statistique et également tenir compte des aspects « terrain » du métier, notamment les limites matérielles et humaines d'investigation au niveau local.

Le développement et l'évaluation de nouvelles méthodes de détection est actuellement en cours à l'InVS. Ces étapes participent à la définition d'un futur système de surveillance du risque de GEA d'origine hydrique en France. Dans ce système, le rôle des investigateurs de terrain sera essentiel car il conditionne la surveillance des facteurs du risque épidémique et augmente les possibilités de prévention.

## Références bibliographiques

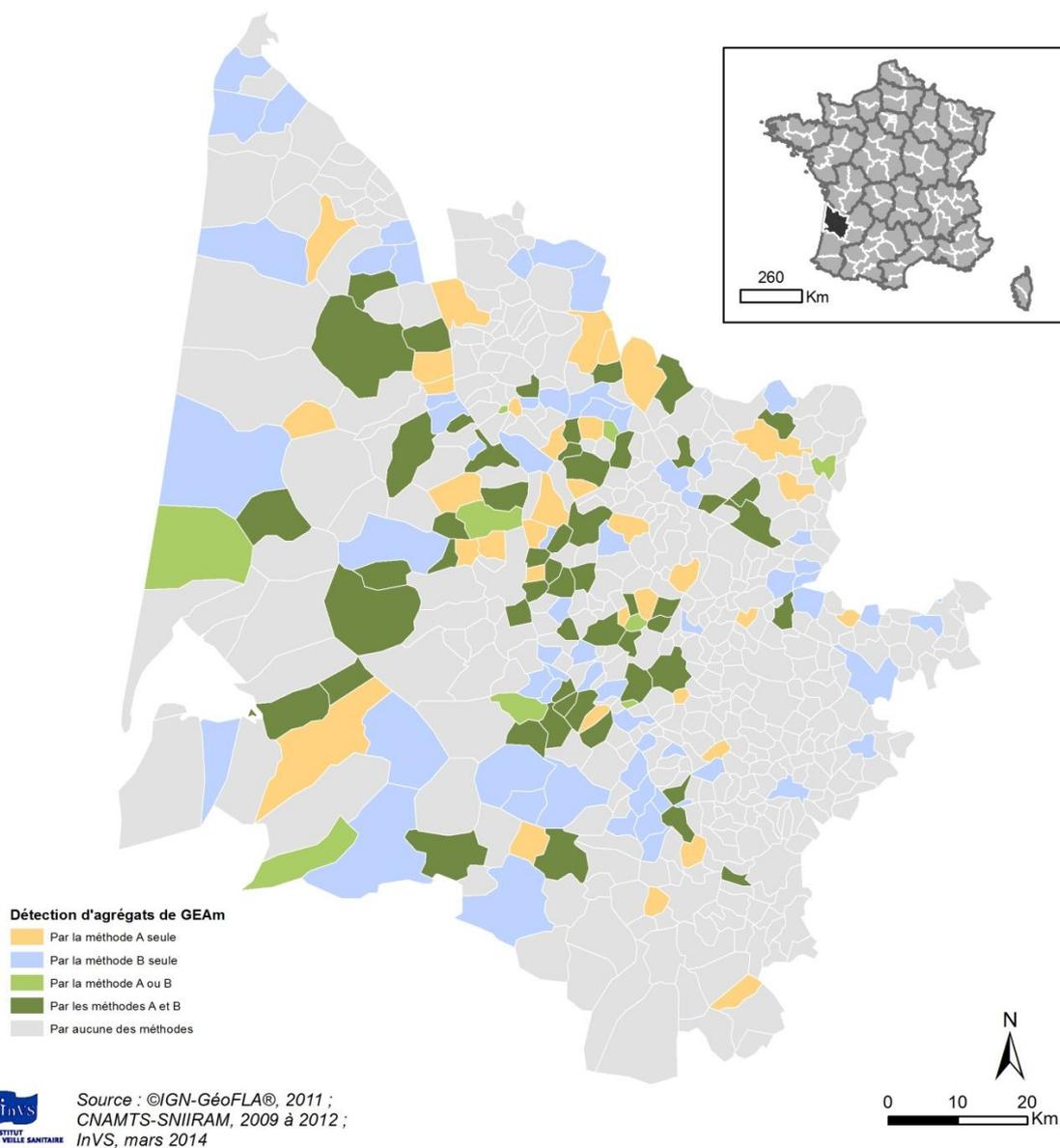
- [1] Beaudreau P, de Valk H, Vaillant V, Mouly D, Ledrans M, Mannschott C, *et al.* Lessons learned from ten investigations of waterborne gastro-enteritis outbreaks, France, 1998-2006. *J Water Health* 2008;6(4):491-503.
- [2] Brunkard JM, Ailes E, Roberts VA, Hill V, Hilborn ED, Craun GF, *et al.* Surveillance for waterborne disease outbreaks associated with drinking water-United States, 2007-2008. *MMWR Surveill Summ* 2011;60(12):38-68.
- [3] Beaudreau P. Surveillance syndromique des gastro-entérites aiguës : une opportunité pour la prévention du risque infectieux attribuable à l'ingestion d'eau du robinet [Syndromic surveillance of acute gastroenteritis: an opportunity for the prevention of the infectious risk attributable to tap water] Université de Rennes 1; 2012. 239 p.
- [4] Tuppin P, de Roquefeuil L, Weill A, Ricordeau P, Merliere Y. French national health insurance information system and the permanent beneficiaries sample. *Rev Epidemiol Santé Publique* 2010;58(4):286-90.
- [5] Bounoure F, Beaudreau P, Mouly D, Skiba M, Lahiani-Skiba M. Syndromic Surveillance of Acute Gastro-Enteritis Based on Drug Consumption, France. *Epidemiol Infect* 2010.
- [6] Majowicz SE, Hall G, Scallan E, Adak GK, Gauci C, Jones TF, *et al.* A common, symptom-based case definition for gastro-enteritis. *Epidemiol Infect* 2008;136(7):886-94.
- [7] Beaudreau P. Déterminants naturels et humains des pollutions fécales accidentelles des petites unités d'alimentation en eau potable, France, 2003-2004 : une étude écologique géographique [Natural and human factors of faecal contamination events of drinking water in small distribution networks, France, 2003-2004: a geographical ecological study]. Saint-Maurice: 2008. 25 p.
- [8] 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.
- [9] Beaudreau P, Bounoure F, Genet M. Estimation de l'incidence des gastro-entérites aiguës à partir des bases de données de l'Assurance Maladie [A method for assessing acute gastro-enteritis incidence from French health insurance databases]. *Pratiques et organisation des Soins* 2006;37(4):313-23.
- [10] Tam CC, Rodrigues LC, O'Brien SJ. The study of infectious intestinal disease in England: what risk factors for presentation to general practice tell us about potential for selection bias in case-control studies of reported cases of diarrhoea. *Int J Epidemiol* 2003;32(1):99-105.
- [11] Delmas G, Gallay A, Espié E, Haeghebaert S, Pihier N, Weill FX, *et al.* Les toxi-infections alimentaires collectives en France entre 1996 et 2005. *BEH* 2006;(51-52):418-22.
- [12] Van Cauteren D, de Valk H, Vaux S, Le Strat Y, Vaillant V. Burden of acute gastro-enteritis and healthcare-seeking behaviour in France: a population-based study. *Epidemiol Infect* 2012;140(4):697-705.
- [13] Smith A, Reacher M, Smerdon W, Adak GK, Nichols G, Chalmers RM. Outbreaks of waterborne infectious intestinal disease in England and Wales, 1992-2003. *Epidemiol Infect* 2006;1-9.

- [14] Haeghebaert S, Le Querrec F, Vaillant V, Delarocque-Astagneau E, Bouvet P. Les toxi-infections alimentaires collectives en France en 1998. BEH 2001 [consulté le 10/04/2001];(15):65-70.
- [15] Beaudeau P, de Valk H, Vaillant V, Mouly D. Détection et investigation des épidémies d'infections liées à l'ingestion d'eau de distribution. Approche intégrée environnementale et épidémiologique. [Detection and investigation of outbreaks of infection related to tapwater. An integrated approach]. Saint-Maurice: Institut de veille sanitaire; 2007. 104 p.
- [16] Rambaud L, Mouly D, Schmitt M, Kerrien F, Beaudeau P. Utilisation des données de remboursement des médicaments de l'Assurance Maladie pour identifier et caractériser une épidémie de gastro-entérites d'origine hydrique, Bourg Saint-Maurice (Arcs 1800), 2006 [Use of drug reimbursements data from French national health insurance to characterize a waterborne outbreak in Bourg Saint-Maurice (Arc 1800), France, 2006]. BEH 2011;(31):339-43.
- [17] Tillier C. Vingt ans d'épidémies d'origine hydrique en Saône-et-Loire. Rapport de la Ddass de Saône-et-Loire. 1-7. 2000.

# Annexes

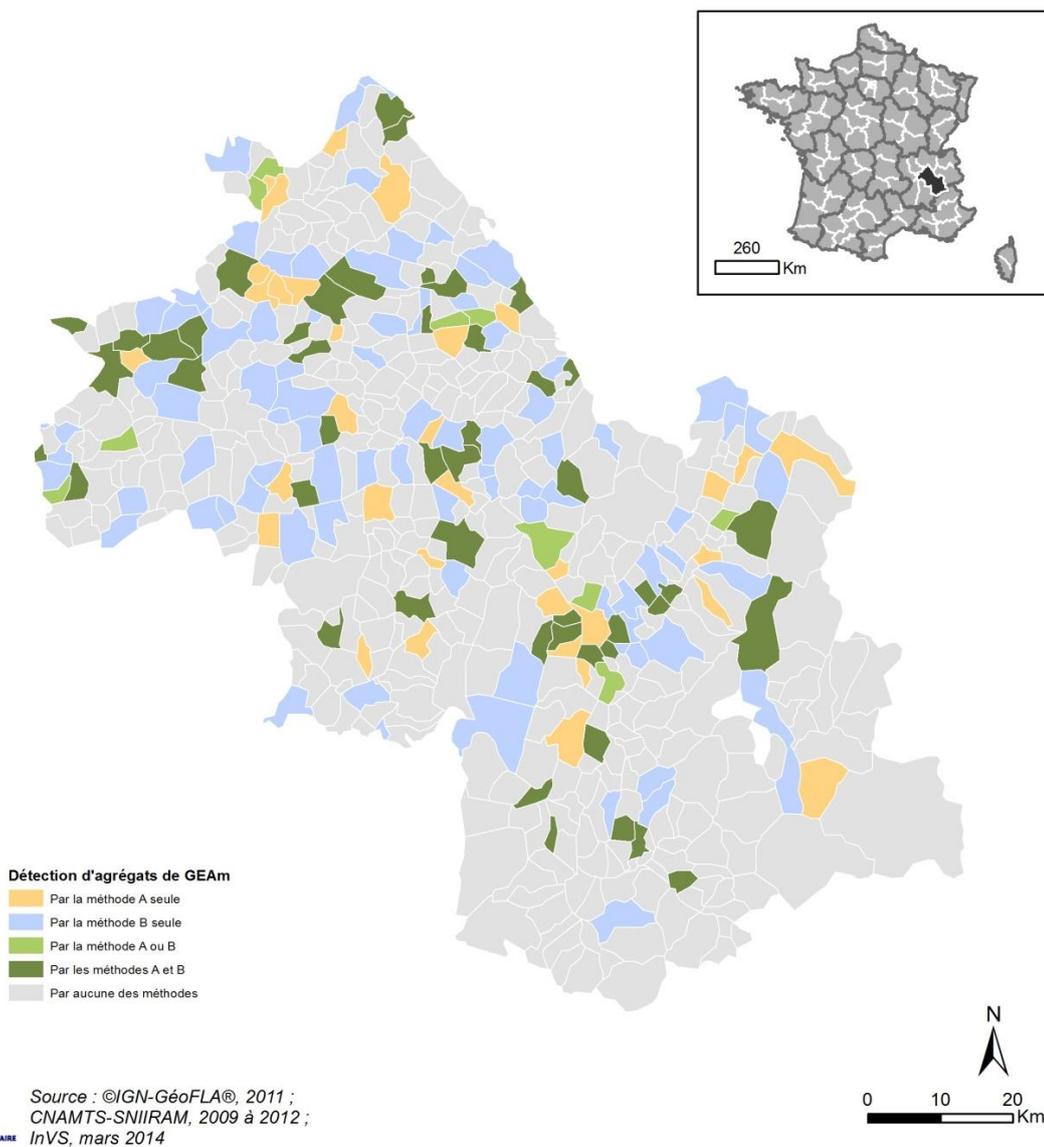
## Annexe A1

### Détection de clusters de gastroentérites médicalisées selon différentes méthodes (A, B, A ou B, A et B), Département de la Gironde (33), 2009 à 2012



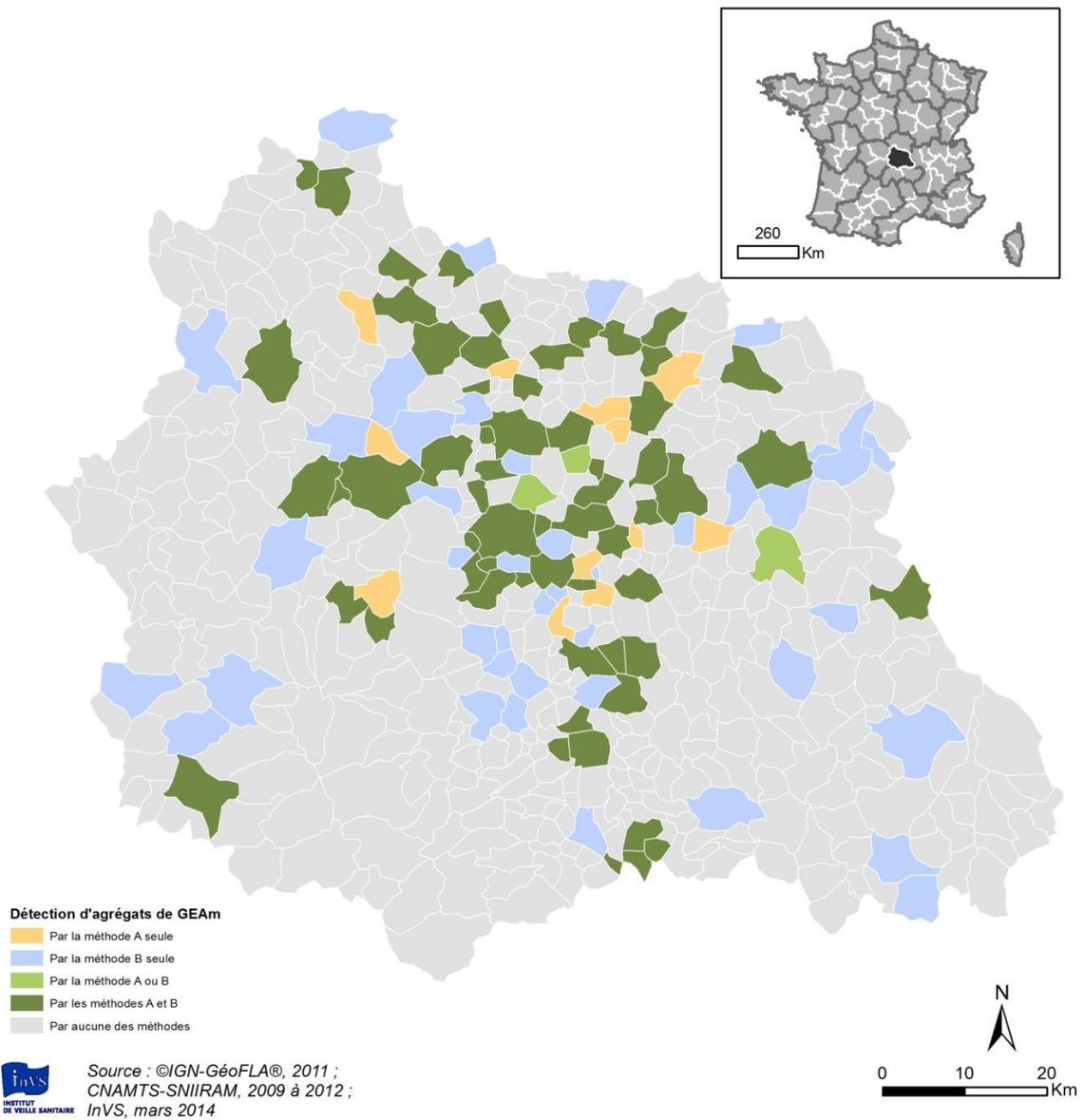
## Annexe A2

### Détection de clusters de gastroentérites médicalisées selon différentes méthodes (A, B, A ou B, A et B), Département de l'Isère (38), 2009 à 2012



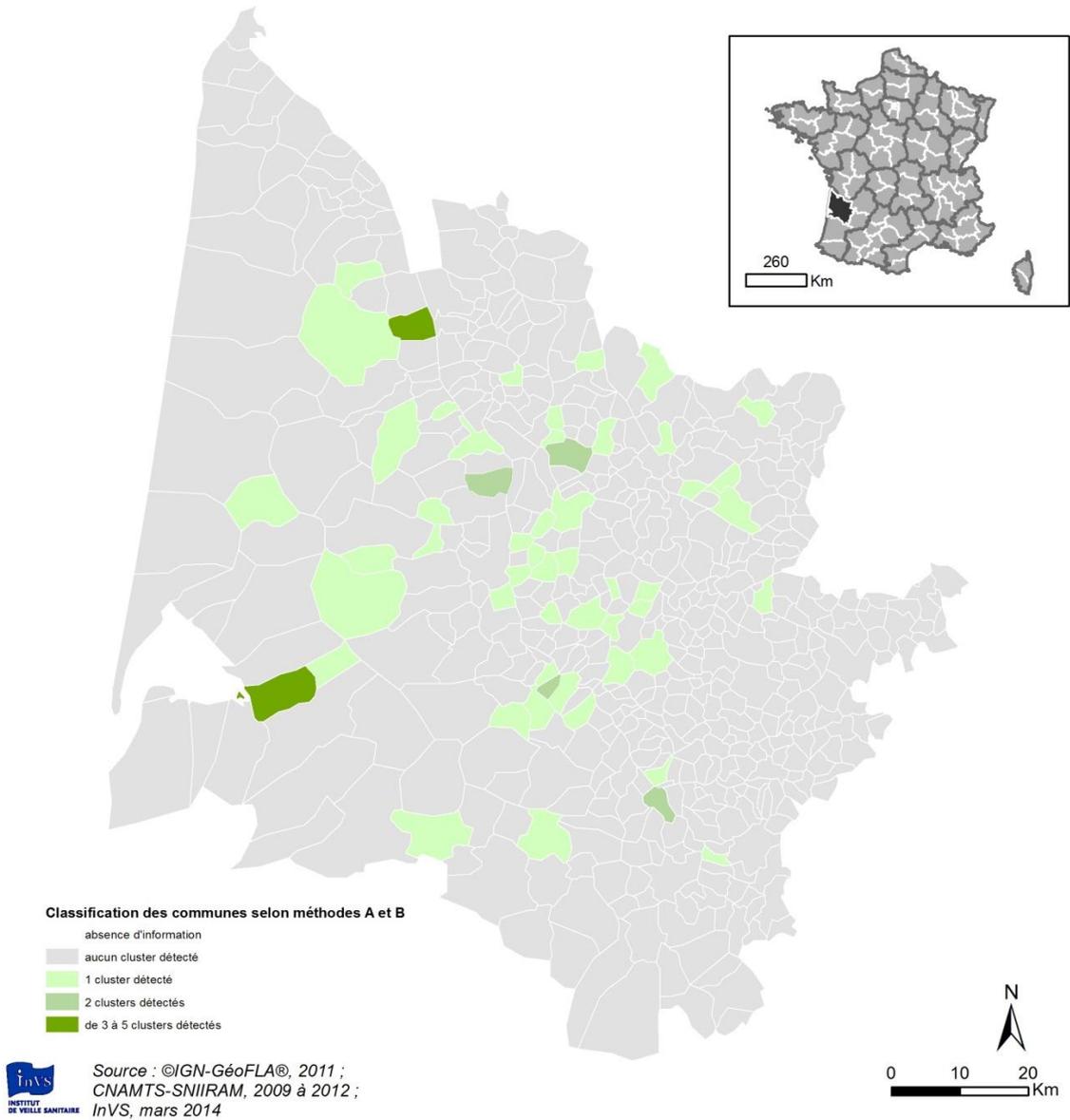
## Annexe A3

### Détection de clusters de gastroentérites médicalisées, selon différentes méthodes (A, B, A ou B, A et B), Département du Puy de Dôme (63), 2009 à 2012



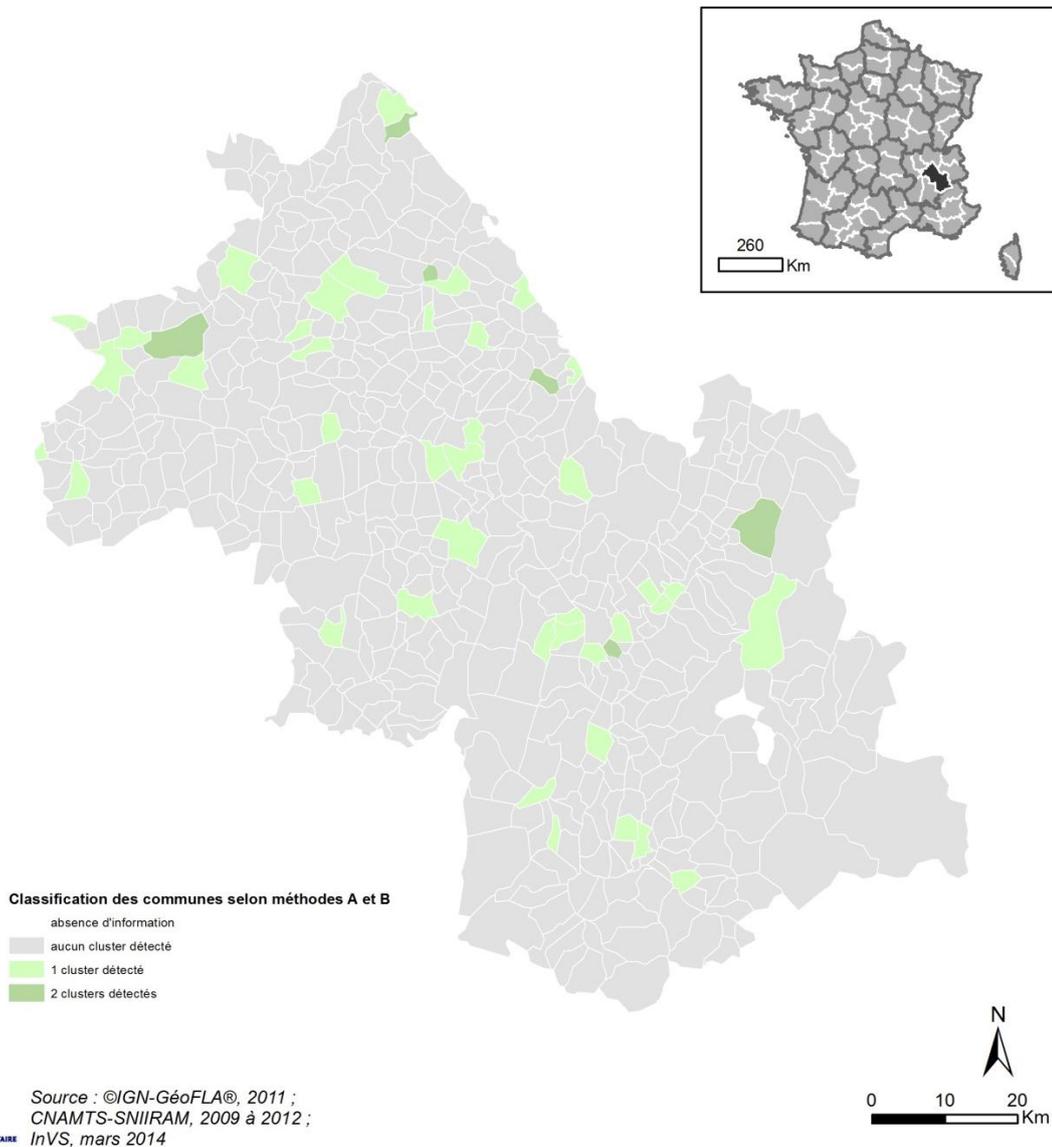
# Annexe B1

## Répartition par commune des clusters de gastroentérites médicalisées détectés par les deux méthodes A et B, Département de la Gironde (33), 2009 à 2012



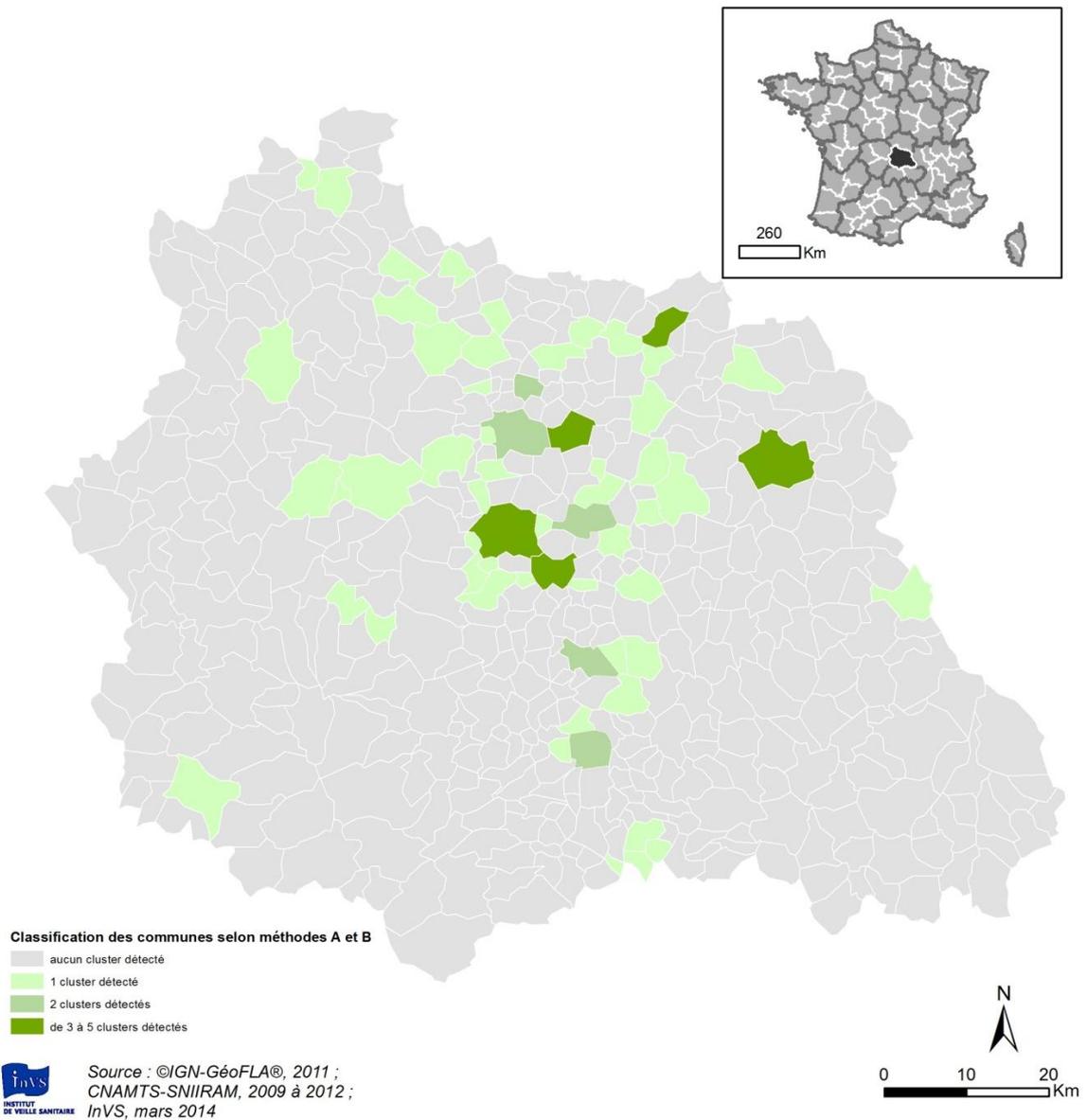
## Annexe B2

### Répartition par commune des clusters de gastroentérites médicalisées détectés par les deux méthodes A et B, Département de l'Isère (38), 2009 à 2012



## Annexe B3

### Répartition par commune des clusters de gastroentérites médicalisées détectés par les deux méthodes A et B, Département du Puy de Dôme (63), 2009 à 2012



## Détection automatisée d'agrégats de cas de gastro-entérites aiguës dans trois départements, France, 2009-2012

### Utilisation pour la prévention des épidémies d'origine hydrique

Une étude pilote a été menée pour appréhender le potentiel des données de remboursement des médicaments de l'Assurance Maladie pour l'identification automatisée (*clustering*) des épidémies d'origine hydrique. Le nombre hebdomadaire de cas de gastroentérite aiguë médicalisée (GEAm) a été dérivé de cette base de données pour les 1 543 communes de 3 départements français à dominante rurale sur la période 2009-2012. Une méthode de clustering, combinant 2 méthodes simples, a été fondée sur la comparaison spatiale des taux d'incidence (méthode A) et sur la comparaison spatiale de l'évolution temporelle de ces taux d'incidence (méthode B) entre la commune ciblée et les communes du département. Pour chaque département, chaque commune a été testée, semaine par semaine, sur toute la période d'étude.

Deux cent dix agrégats hebdomadaires impliquant plus de 5 cas de GEAm ont été identifiés par cette méthode à partir de critères de détection tirés de la pratique des investigations de terrain. Ces agrégats forment 193 agrégats consolidés de 1 à 3 semaines. Une commune sur dix a éprouvé un agrégat et moins de 2 sur 100 plusieurs. L'inférence à la France entière indique que 1 000 agrégats (soit 30 000 cas de GEAm au total) pourraient être détectés annuellement.

L'impact représenté par les agrégats ne comprend qu'une faible part du total des cas de GEAm d'origine hydrique. Plus que de renseigner l'impact infectieux de l'eau, l'objectif de ce *clustering* est de stimuler les investigations de terrain, les orienter vers les situations les plus à risque en coordination avec la mise en place des Water Safety Plans, et alimenter la surveillance des facteurs de risque au niveau national. Cette méthode apparaît perfectible. Une inter-comparaison avec d'autres méthodes doit être menée par ailleurs, avant le choix définitif de la méthode qu'il conviendra de mettre en œuvre au niveau national.

**Mots clés :** base de données médico-tarifaires, clustering, eau distribuée, épidémie, France, gastroentérite, surveillance épidémiologique

## Automated clustering of gastroenteritis cases in three French departments, 2009-2012

### Use for the prevention of waterborne outbreaks

*A pilot study was carried out to assess the potential of the drug reimbursement database from the French national health insurance system to support automated clustering for the detection of waterborne outbreaks. The weekly counts of medical prescriptions for acute gastroenteritis (AGE) were drawn from this database for the 1,543 communes of three rural districts in 2009-2012. We developed a clustering method combining two simple methods. One was based on the spatial comparison of the incidence rate (method A), and the other the spatial comparison of the temporal evolution of the incidence rate (method B) between the targeted commune and the district's communes. We thus screened the three districts. For each one, we tested each commune of the district, week by week through the study period.*

*We found 210 "week.commune" clusters of medicalized AGE (mAGE) cases meeting the practice-based inclusion criteria for both methods A and B, representing 193 consolidated clusters that lasted from 1 to 3 weeks. Ten per cent of the communes were affected by at least one cluster and 1.5% by several. The inference to whole France led to an assessment of >1,000 clusters per year, corresponding to a total of 30,000 mAGE cases.*

*We estimated the burden due to mAGE clusters within the total burden of AGE, and concluded that clustering only highlights a small part of the burden of disease attributable to tap water. More than inform the burden of disease, cluster monitoring mainly aims to trigger and guide field investigations of water systems in communes at greater risk. Field investigations could help guide the implementation of the water safety plans and feed at national level the surveillance of risk factors to support prevention. Finally, ways to improve the method are discussed. The benchmarking of our method along with other available methods is expected before the choice of the method to be implemented at national level.*

Citation suggérée :

Rambaud L, Galey C, Guillet A, Corso M, Van Cauteren D, Beaudreau P. Détection automatisée d'agrégats de cas de gastro-entérites aiguës dans trois départements, France, 2009-2012. Utilisation pour la prévention des épidémies d'origine hydrique. Saint-Maurice : Institut de veille sanitaire ; 2014. 34 p. Disponible à partir de l'URL : <http://www.invs.sante.fr>

INSTITUT DE VEILLE SANITAIRE

12 rue du Val d'Osne

94415 Saint-Maurice Cedex France

Tél. : 33 (0)1 41 79 67 00

Fax : 33 (0)1 41 79 67 67

[www.invs.sante.fr](http://www.invs.sante.fr)

ISSN: 1958-9719

ISBN-NET: 979-10-289-0084-7

Réalisé par Service communication – InVS

Dépôt légal : octobre 2014