Waterborne disease outbreak detection: an integrated approach using health administrative databases

S. Coly, N. Vincent, E. Vaissiere, M. Charras-Garrido, A. Gallay, C. Ducrot and D. Mouly

ABSTRACT

Hundreds of waterborne disease outbreaks (WBDO) of acute gastroenteritis (AGI) due to contaminated tap water are reported in developed countries each year. Such outbreaks are probably under-detected. The aim of our study was to develop an integrated approach to detect and study clusters of AGI in geographical areas with homogeneous exposure to drinking water. Data for the number of AGI cases are available at the municipality level while exposure to tap water depends on drinking water networks (DWN). These two geographical units do not systematically overlap. This study proposed to develop an algorithm which would match the most relevant grouping of municipalities with a specific DWN, in order that tap water exposure can be taken into account when investigating future disease outbreaks. A space-time detection method was applied to the grouping of municipalities. Seven hundred and fourteen new geographical areas (groupings of municipalities) were obtained compared with the 1,310 municipalities and the 1,706 DWN. Eleven potential WBDO were identified in these groupings of municipalities. For ten of them, additional environmental investigations identified at least one event that could have caused microbiological contamination of DWN in the days previous to the occurrence of a reported WBDO.

Key words | drinking water, ecological study, public health surveillance, space-time detection, waterborne disease outbreaks

S. Coly
M. Charras-Garrido
C. Ducrot
INRA, UR346 – Unité d'Épidémiologie Animale,
Centre de recherche de Clermont-Ferrand,
63122 Saint Genès Champanelle,
France

N. Vincent
E. Vaissiere
A. Gallay
D. Mouly (corresponding author)
French National Public Health Agency,
12 rue du Val d'Osne,
94 415 Saint-Maurice Cedex,
France
E-mail: *damien.mouly@santepubliquefrance.fr*

ABBREVIATIONS

AGI	Acute gastrointestinal infection
DWN	Drinking water network
GP	General practitioner
ANSP	Agence Nationale de santé publique (French National Public Health Agency)
SISE-eaux	Système d'Information en Santé-Environnement sur les Eaux d'alimentation (Information system on environmental health – drinking water supply)
SNIIRAM	Système national d'information inter-régimes de l'Assurance maladie (French National Health Insurance Information System)
WBDO	Waterborne disease outbreak

doi: 10.2166/wh.2017.273

INTRODUCTION

Waterborne disease outbreaks (WBDO) are a public health concern in developed countries because of the large proportion of people potentially affected when contamination of drinking water occurs (Hrudey & Hrudey 2004; Beaudeau *et al.* 2008; Craun *et al.* 2010). To date, detection of these events has mainly occurred through general practitioners' (GPs) reporting of clusters of acute gastrointestinal infection (AGI) to health authorities. The absence of a designated surveillance system suggests therefore that the number of WBDO is probably underestimated in France. In public health terms, increasing the detection of infections caused by contaminated drinking water contributes to improving the following factors: knowledge of risk factors, identification of high risk drinking water networks (DWN), and development of appropriate preventive measures. In this context, the French National Public Health agency (ANSP) is exploring the possibility of using the health administrative databases from the French Health Insurance System to develop a national automated detection system for WBDO.

Searching for a link between a health indicator (e.g., a WBDO) and associated environmental exposure factors (e.g., drinking water consumption), is a frequent subject of study in the field of epidemiological surveillance (Chaput *et al.* 2002; Klassen *et al.* 2005).

To date, most related studies have considered health and environmental data separately (Mostashari *et al.* 2003; Hayran 2004; Osei & Duker 2008), by first attempting to detect spatial or spatiotemporal areas in which clusters of cases occur, and mapping environmental exposure factors. Then, the locations of case clusters and factors linked to the environmental area of exposure are compared (Fukuda *et al.* 2005). Other studies in the literature have successfully considered both health concerns and environmental factors together. Most of these have taken a common approach (Patil & Taillie 2004) whereby a statistical method is first applied to detect the occurrence of a cluster of cases. Then, a validation test of the detected clusters is performed, followed by identification a posteriori of the environmental factors related to each cluster. However, results of tests to validate and explain detected clusters have not been conclusive in most of these studies (D'Aignaux *et al.* 2002; Odoi *et al.* 2004).

In our study, within the framework of the detection of WBDO, the environmental exposure factor considered was the DWN (or distribution zone). Our hypothesis was that a DWN delivers water of homogenous microbiological quality to consumers, i.e., any individuals connected to the same DWN are similar from the point of view of water quality. For this reason, DWN was considered as the environmental factor of interest for the detection of WBDO. Therefore, the area covered by a DWN was considered as the spatial unit of interest to study clusters of WBDO. The health data considered in the present study came from the French National Health Insurance Information System (SNIIRAM: Système national d'information inter-régimes de l'Assurance maladie).

The aim of the study presented in this article was to develop an integrated approach to detect and study clusters of AGI in geographical areas with homogeneous exposure to drinking water, for which data on the human population and cases of AGI were both available.

This newly developed approach was tested on real AGI data in France.

METHODS

The integrated approach we used needed to manage the absence of systematic overlapping between exposure data area (DWN) and cases of AGI data area (municipality). This was a two-step approach, as follows: first, the creation of new geographical units taking into account a priori drinking water exposure and aggregation of cases of AGI; second, the application of a space-time detection method of clusters of AGI in these geographical units (Kulldorff *et al.* 2005). The method is briefly detailed in a following section.

Three types of data were used: health data, geographical and population data, and environmental data.

Health data and case definition

The health indicator used in our study for the detection of WBDO was cases of AGI following a medical visit by a GP.

In 2011, an algorithm was specifically developed in France to identify AGI cases by using data on reimbursement for payment of prescribed drugs from the SNIIRAM database (Bounoure *et al.* 2011). The SNIIRAM aims at evaluating beneficiaries' healthcare consumption and associated expenditures. It covers more than 98% of the French population and records all reimbursements to patients for out-of-pocket medical procedures, medications, and payments to professionals for consultations (Tuppin *et al.* 2010). AGI medications which are reimbursable, prescribed by a GP and dispensed in a pharmacy are included in this database. The identification of AGI cases required two consecutive steps: (i) data extraction from the SNIIRAM database and (ii) use of the AGI algorithm developed by Bounoure *et al.* (2011) during a pharmacy-based survey to select AGI cases. The criterion for the first step was as follows: reimbursement for at least one prescribed target drug used to treat AGI (antiemetic drugs – ATC classification: A04A, A03F; antidiarrhea drugs – A07X, A07D; intestinal adsorbent drugs – A07B, A02X and oral rehydration salts). The criteria for the AGI discriminative algorithm were as follows: a delay of <24 hours between the prescription and delivery of drugs, the number of different AGI-specific drugs prescribed, treatment duration (less than 8 days), and the co-prescription of non-AGI specific drugs (e.g., anti-cancer drugs). In the study by Bounoure *et al.* (2011), the sensitivity and specificity of the AGI algorithm were estimated by the ability of the algorithm to provide a conclusive answer about the existence or not of an AGI case compared with the diagnosis verbally reported by patients participating in the pharmacy-based survey ($n = 557$ individuals). Both indicators reached almost 90% (Bounoure *et al.* 2011).

Data on age, gender, date of consultation, and municipality of residence were available for each case of AGI and cases were aggregated by municipality of residence.

The AGI database used for analysis contained the number of new cases for each day and for each municipality of residence.

Geographical data

The data describing the municipalities' coordinates came from the national geographic institute (Institut National de l'Information Géographique et Forestière 2013).

Environmental and population data

To take account of the environmental exposure factor (DWN), we used the Information system on environmental health – drinking water supply (SISE-eaux: 'Système d'Information en Santé-Environnement sur les Eaux d'alimentation') managed by the French Ministry of Health. This database contains the list of all DWN in France and the list of the municipalities served. For each DWN, technical information about installations is also available (e.g., water treatment plant, tank). Moreover, population data correspond to the number of people served by each DWN, the number of inhabitants in each municipality, and the number of people served at the overlap of DWN and served municipalities.

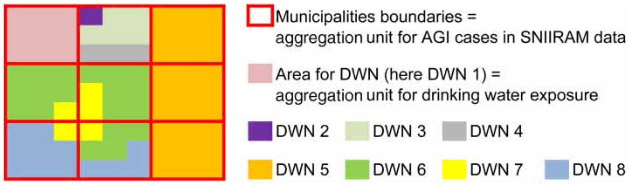

Municipalities boundaries = aggregation unit for AGI cases in SNIIRAM data
Area for DWN (here DWN 1) = aggregation unit for drinking water exposure
DWN 2
DWN 3
DWN 4
DWN 5
DWN 6
DWN 7
DWN 8

Figure 1 | Population number for DWN, municipalities and intersection DWN/municipalities.

Data extraction was performed by selecting the following variables: French county number, DWN code and name, zip codes and names of municipalities served by DWN, population numbers for DWN (Di in Figure 1), number of inhabitants of each municipality (Mj in Figure 1), and population number at the overlap of DWN and municipalities served (e.g., Xij in Figure 1).

Study area and period

The study area was an administrative region in the center of France (Auvergne). This area was chosen because of the existence of an environmental and sanitary signal associated with tap water, specifically, the occurrence of frequent microbiological contaminations of DWN and of WBDO (Mouly *et al.* 2016).

Health data were collected for the period between January 1, 2009 and December 31, 2012.

Description of an algorithm used to define area with homogeneous tap water exposure in WBDO detection system

The geographic areas for aggregation of cases of AGI (municipality level) and for exposure to drinking water (DWN area) do not always overlap. There are four configurations to represent the correspondence between DWN geographical limits and municipality boundaries in France (Figure 2), as follows. (i) 1 municipality for 1 DWN (e.g., DWN 1 in Figure 2). This is the perfect overlapping configuration. The aggregated unit of health data (cases of AGI) corresponds exactly to a single aggregated unit of tap water exposure (i.e., one single DWN). (ii) 1 municipality = n DWN indicates that the population of the municipality is served by different DWN so the population of the same

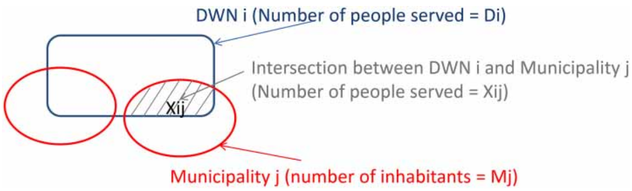
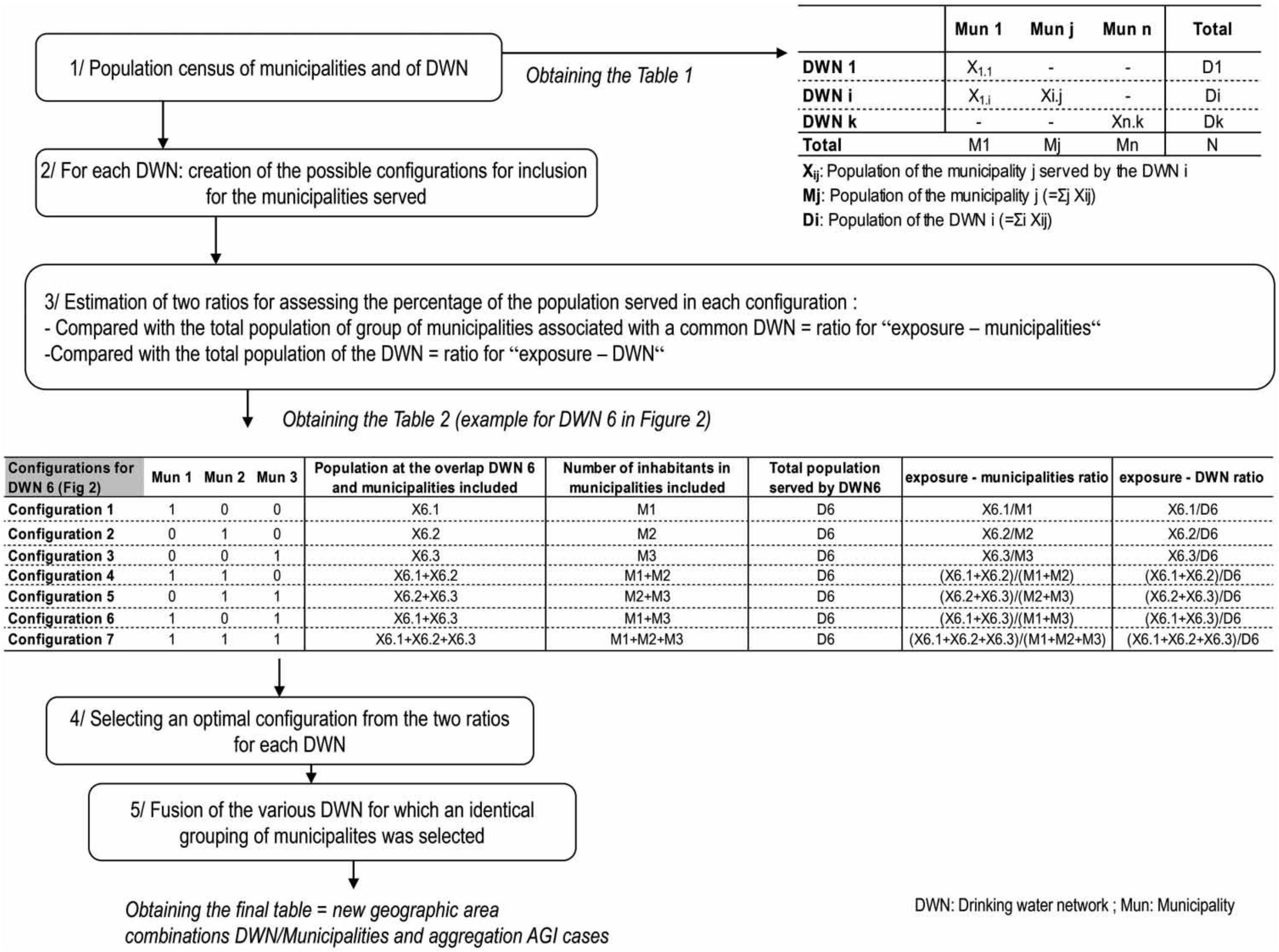

DWN i (Number of people served = Di)
Intersection between DWN i and Municipality j
(Number of people served = Xij)
Xij
Municipality j (number of inhabitants = Mj)

Figure 2 | Several possible configurations for overlapping between DWN area and municipalities' boundaries.

municipality may be exposed to heterogeneous tap water quality (e.g., DWN 2–4). (iii) m municipalities = 1 DWN indicates that only one DWN serves several municipalities (e.g. DWN 5). The whole population of these municipalities drinks water of the same quality. (iv) m municipalities = n DWN is the most complicated configuration, because there is no direct relationship whatsoever between exposure by aggregation unit (i.e., DWN) and health data aggregation unit (i.e., municipalities) (e.g., DWN 6–8).

Based on the different possible configurations between DWN and municipalities, the algorithm contained the following phases (Figure 3):

1. *Listing the municipalities served by each DWN and evaluating the corresponding population size* (step 1, Figure 3). For each DWN, all the municipalities partially or globally served were recorded. The following variables were specified:
   - total number of people served by the DWN (Di, Figure 1);
   - total number of people living in each municipality (Mj, Figure 1);



1/ Population census of municipalities and of DWN

Obtaining the Table 1

	Mun 1	Mun j	Mun n	Total
DWN 1	$X_{1.1}$	-	-	D1
DWN i	$X_{1.i}$	Xi.j	-	Di
DWN k	-	-	Xn.k	Dk
Total	M1	Mj	Mn	N

$\mathbf{X_{ij}}$: Population of the municipality j served by the DWN i
Mj: Population of the municipality j (=Σj Xij)
Di: Population of the DWN i (=Σi Xij)

2/ For each DWN: creation of the possible configurations for inclusion for the municipalities served

3/ Estimation of two ratios for assessing the percentage of the population served in each configuration :
- Compared with the total population of group of municipalities associated with a common DWN = ratio for "exposure – municipalities"
-Compared with the total population of the DWN = ratio for "exposure – DWN"

Obtaining the Table 2 (example for DWN 6 in Figure 2)

Configurations for DWN 6 (Fig 2)	Mun 1	Mun 2	Mun 3	Population at the overlap DWN 6 and municipalities included	Number of inhabitants in municipalities included	Total population served by DWN6	exposure - municipalities ratio	exposure - DWN ratio
Configuration 1	1	0	0	X6.1	M1	D6	X6.1/M1	X6.1/D6
Configuration 2	0	1	0	X6.2	M2	D6	X6.2/M2	X6.2/D6
Configuration 3	0	0	1	X6.3	M3	D6	X6.3/M3	X6.3/D6
Configuration 4	1	1	0	X6.1+X6.2	M1+M2	D6	(X6.1+X6.2)/(M1+M2)	(X6.1+X6.2)/D6
Configuration 5	0	1	1	X6.2+X6.3	M2+M3	D6	(X6.2+X6.3)/(M2+M3)	(X6.2+X6.3)/D6
Configuration 6	1	0	1	X6.1+X6.3	M1+M3	D6	(X6.1+X6.3)/(M1+M3)	(X6.1+X6.3)/D6
Configuration 7	1	1	1	X6.1+X6.2+X6.3	M1+M2+M3	D6	(X6.1+X6.2+X6.3)/(M1+M2+M3)	(X6.1+X6.2+X6.3)/D6

4/ Selecting an optimal configuration from the two ratios for each DWN

5/ Fusion of the various DWN for which an identical grouping of municipalites was selected

Obtaining the final table = new geographic area combinations DWN/Municipalities and aggregation AGI cases

DWN: Drinking water network ; Mun: Municipality

Figure 3 | Pattern of new geographical areas based on population size data in DWN and municipalities.

- total number of people served by the DWN in each municipality (Xij, Figure 1).

2. *Listing all the possible configurations between DWN and municipalities* (step 2, Figure 3). Two situations were distinguished:

   1st situation: All the municipalities were served by only one DWN (configurations '1 municipality = 1 DWN' and 'm municipalities = 1 DWN'). In this case it was supposed that all the municipalities were fully exposed when pollution occurs. Consequently, they were all systematically included in the possible configurations associated with this DWN.

   2nd situation: Municipalities were partially served by a DWN (configurations '1 municipality = n DWN' and 'm municipalities = n DWN'). In this case, including or excluding this municipality in the grouping which best matched the DWN had to be decided. If k is the number of municipalities partially served by the DWN, 2^{k-1} is the possible number of groupings of municipalities which must be considered for inclusion or not.

3. *Computing indicators for each configuration* (step 3, Figure 3).

   For each DWN, two indicators were built for each possible grouping of municipalities:

   - The '**exposure-municipalities ratio**'. This is the ratio of the population size served by a DWNi in a grouping of municipalities (ΣiXij) to the total population of all the concerned municipalities (Σ Mj). The higher the ratio, the greater the capability of the statistical methods employed to detect low-intensity WBDO (i.e., higher power of detection).
   - The '**exposure-DWN ratio**'. This is the ratio of the population size served by a DWNi in a grouping of municipalities (ΣiXij) to the total population served by the DWN (Di). The higher this ratio, the stronger the likelihood that a potential outbreak of AGI in this grouping is due to exposure to contaminated drinking water (likelihood of a WBDO).

4. *Selection rule based on these indicators to determine the final configuration for each DWN* (step 4, Figure 3).

   The selection of the optimized configuration (grouping of municipalities) for each DWN was made by minimizing the Euclidian distance between (exposure-municipalities ratio, exposure-DWN ratio) and (1, 1): $M = (1 - \textit{exposure-municipalities ratio})^2 + (1 - \textit{exposure-DWN ratio})^2$

   The grouping of municipalities chosen was the one associated with the smallest value of M.

5. *Merging the DWN corresponding to the same grouping of municipalities* (step 5, Figure 3). After the four previous steps, certain DWN corresponded to the same grouping of municipalities. We merged these to avoid any problems of repetition.

Cases of AGI were then aggregated over the new geographical area created by algorithm before the cluster detection process.

The algorithm was implemented using R software (versions 2.14 and 2.15).

Space-time detection of cluster of AGI

Several published methods for cluster detection are available in the literature (Mostashari *et al.* 2003; Kulldorff *et al.* 2005; Takahashi *et al.* 2008; Assuncao & Correat 2009; Cucala 2009). We selected the space-time detection method, developed by Kulldorff *et al.* (2005). This method has been widely used in the literature and would appear to be a reference method for cluster detection in epidemiological surveillance (Heffernan *et al.* 2004; Balter *et al.* 2005; Assuncao & Correat 2009). Moreover, it presents several interesting criteria for its use for WBDO detection including: the consideration of seasonality during winter epidemics of AGI, the possibility to use covariates (for example, the day of the week and public holidays), the consideration of the multiple comparison problem, and the simplicity of the method's application with SatScan software.

Space-time permutation of Kulldorff's method (Kulldorff *et al.* 1997, 2005) allows areas with excess cases of AGI to be identified in terms of space and time. Applying the method to the algorithm-created geographical units (see previous section) consists of performing a scan of the whole study area, by moving a sliding window located successively at the central point of each geographic unit. Each window is compared with the outer window (which constitutes the entire geographic area under study). For space-time detection, the cylindrical window then travels in time and space so that all geographic units, sizes and durations are successively considered. This results in a great number of windows, and each is a candidate for an AGI cluster. A cluster is detected when the number of cases of AGI within the window is significantly higher than that outside this window. Statistical testing is based on the likelihood ratio, i.e., the ratio of the likelihood calculated under the alternative hypothesis (the risk within the window is greater than that outside), and the calculated likelihood in the null hypothesis of equal risks. The window with the highest likelihood ratio defines the most likely cluster, i.e., the cluster least likely to occur by chance. SatScan v9.3 software was used to implement the Kulldorff method. The following parameters were defined: time aggregation unit for AGI cases (day), aggregation duration for analysis (day), analysis type (retrospective space-time analysis with a space-time permutation model), and finally, type of inference (Monte Carlo inference with 999 replications).

The analysis was performed for the study area each year from January 1, 2009 to December 31, 2012.

Selection of clusters of AGI and validation of their waterborne origin

The clusters obtained were analyzed to select those whose characteristics most reflected the characteristics found for a given WBDO, according to epidemiological knowledge already available regarding that WBDO. Criteria for the selection of clusters were: duration of the signal for more than 6 days, size of the outbreak with more than 10 excess AGI cases, ratio between observed and expected cases of AGI higher than 3, and a *p*-value lower than 0.05.

Finally, the selected clusters were analyzed with the local health authorities to investigate whether specific environmental factors could have pointed to a microbiological contamination of the targeted DWN in the days before clusters appeared. These factors were as follows: results of sanitary control on fecal indicators (*Escherichia coli* and fecal *streptococci*), heavy rains, an incident in the water treatment plant or in the DWN, cessation of disinfection. Furthermore, we checked for the existence of WBDO notification to authorities at the time of the occurrence.

Selected clusters were described using several epidemiological and environmental parameters. The former included the starting date and the duration of the period associated with the cluster, the number of observed and expected cases associated with the cluster, the observed-expected case ratio, the AGI case attack rate (estimated by the ratio between the number of observed cases and the size of population). Environmental parameters included checking for the existence of microbiological pollution during the cluster duration, the percentage of non-microbiological compliance with Ministry of Health fecal indicators during the study period (2009–2012), and the existence of other environmental risk factors (e.g., heavy rain) or a technical incident in the drinking water treatment or distribution system (e.g., water pipe breakages).

RESULTS

Description of configurations of inclusion of municipalities and DWN

The region of Auvergne contains 1,343,964 people living in 1,310 municipalities. The biggest municipality (regional capital, Clermont-Ferrand) contains 139,000 inhabitants. Fifteen municipalities have more than 10,000 inhabitants each, and 81% percent of municipalities have less than 1,000 inhabitants, accounting for 27% of the global regional population.

In Auvergne, 543 of the region's 1,706 DWN serve 20 people or fewer. Indeed, DWN serving 100 people or fewer account for the majority of DWN (62%), but serve only 2.5% of the whole population. Combined, the 10.6% of DWN which each serve more than 1,000 people serve 86.6% of the global population. Only four DWN in Auvergne each serve more than 30,000 people.

The four different matching configurations of DWN and municipalities for our study area are summarized in Table 1 and Figure 4.

Description of the new areas obtained by the algorithm

After applying the algorithm, 714 new geographical areas (Figure 5) were created which grouped together the 1,310

Table 1 | Configurations of inclusion of municipalities and DWN in Auvergne, including number of corresponding municipalities and population size

	Configurations of inclusion municipalities/DWN				
	1 Municipality = 1 DWN	**m Municipalities = 1 DWN**	**1 Municipality = n DWN**	**m Municipalities = n DWN**	**Total**
Municipalities					
N	114	659	241	296	1,310
Percentage	8.7%	50.3%	18.4%	22.6%	100.0%
Population					
N (inhabitants)	151,447	524,630	293,074	374,813	1,343,964
Percentage	11.3%	39.0%	21.8%	27.9%	100.0%

DWN, Drinking water network.

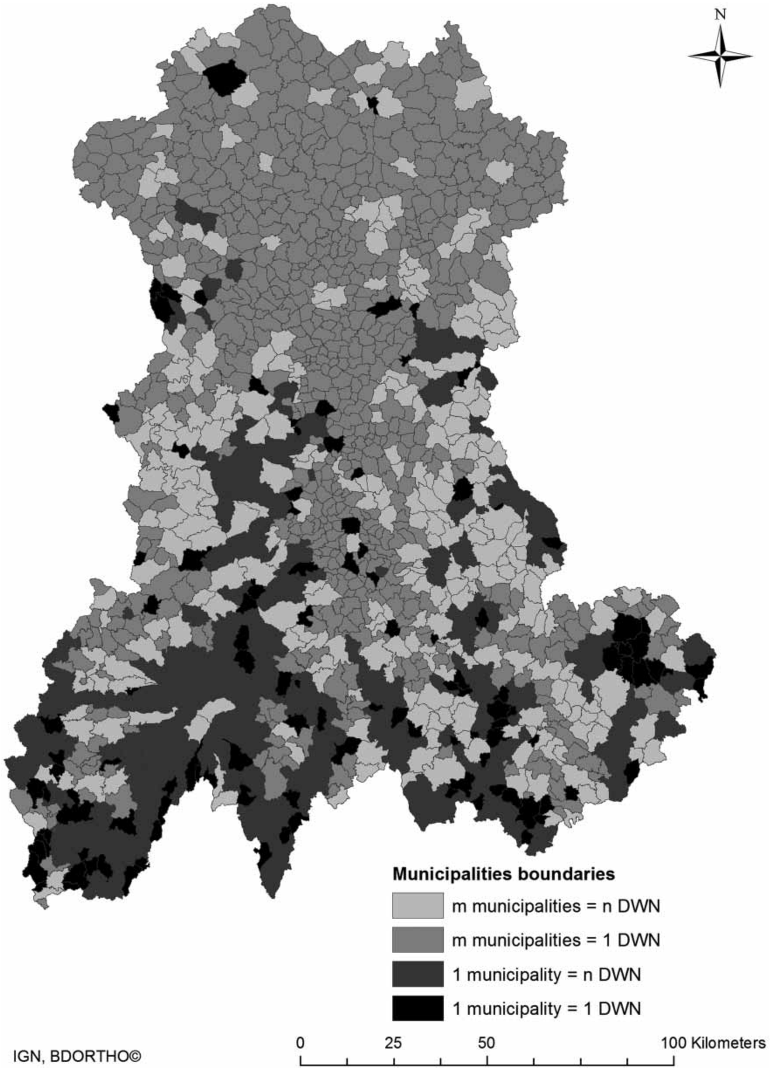


Figure 4 | Map of the configurations of inclusion of municipalities and DWN in Auvergne. *Source:* Sise-Eaux, Ministère chargé de la santé; DWN, Drinking Water Network.

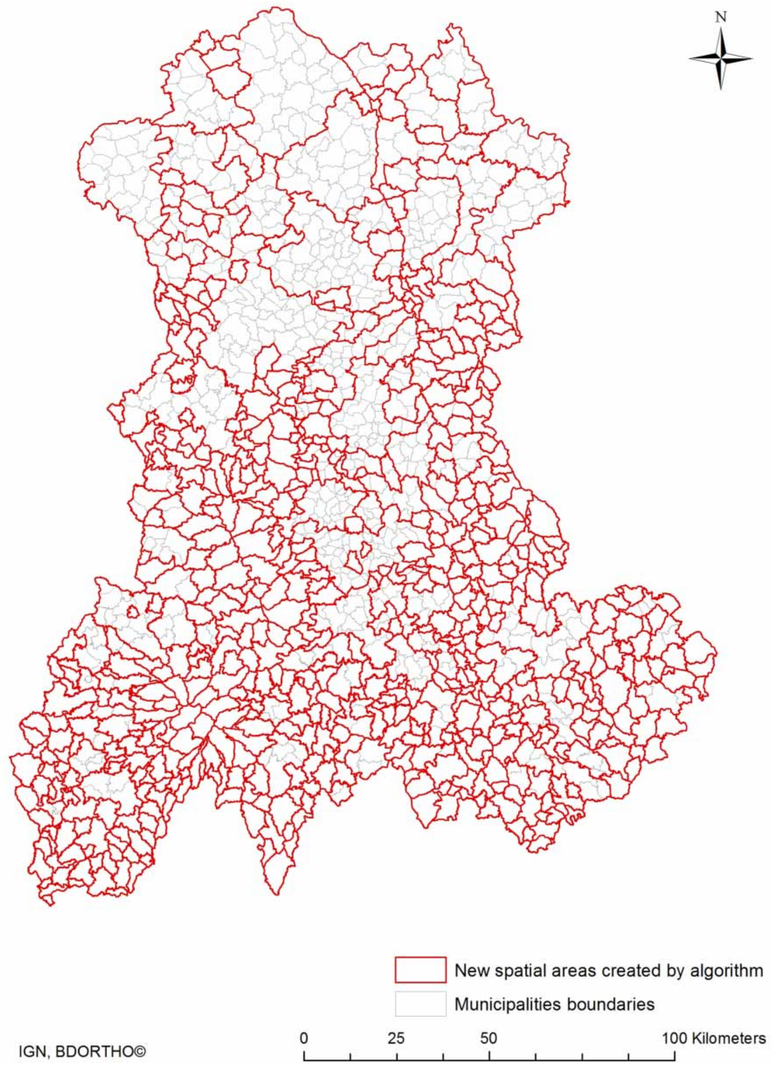

N
🟥 New spatial areas created by algorithm
□ Municipalities boundaries
0
25
50
100 Kilometers
IGN, BDORTHO©

Figure 5 | Spatial delimitation of the new geographical area.

municipalities and 1,706 DWN in Auvergne. The average population size of these areas was 1,891 people.

Most of the new areas contained only one municipality ($n = 573$, 80%). However, 12% were associated with at least three municipalities, accounting for 53% of all the municipalities. These areas were much more concentrated in lowland areas (topographic data not presented). Only 3% of the new areas contained at least ten DWN.

Finally, all municipalities were included at least once in the composition of the resulting new areas. Approximately 91% of the municipalities were associated with only one new area, 9% with at least two areas. Only one municipality was included in four new areas.

Description of clusters

Among all the detected clusters (50 clusters with $p < 0.05$), 11 were consistent with possible WBDO according to the selection criteria above (Table 2 and Figure 6). The impacted grouping of municipalities defined by the algorithm numbered between 500 and 5,000 inhabitants each. Between 20 and 60 cases of AGI were involved in each cluster. The medication rate in the impacted population was approximately 1.4% (median) and varied between 0.7% and 4.8%. The total duration of cumulated WBDO for all the clusters was 177 days, and the longest cluster duration was 35 days. For two of the 11 selected

Table 2 | Description of the 11 clusters of AGI most probably related to contamination of DWN, Auvergne region, 2009–2012

Cluster ID	Year	Area ID	Number of municipalities	Number of DWN	Population served (inhabitants)	Start date	Duration (days)	Observed cases of AGI	Expected cases of AGI	Obs/Exp	Medication rate in population[a]	Microbiological pollution during cluster	% of non-microbiological compliance[b]	Other environmental factors	Notification of WBDO to the health authority	*p*-value
1	2009	707	1	2	4,910	11/26/09	7	67	13.9	4.8	1.4%	No	0.0%	YES	NO	1.0×10^{-17}
5	2010	385	1	3	1,563	03/19/10	19	33	9.59	3.4	2.1%	No	12.2%	YES	NO	6.0×10^{-4}
4	2010	155	1	1	501	03/31/10	12	24	3.25	7.4	4.8%	No	7.1%	YES	NO	2.4×10^{-8}
2	2010	638	1	3	2,650	06/21/10	7	42	4.8	8.8	1.6%	Yes[c]	1.5%	YES	YES	1.0×10^{-17}
6	2010	207	1	4	1,549	08/16/10	35	21	5.29	4.0	1.4%	No	8.5%	NA	NO	4.8×10^{-02}
3	2010	88	12	1	5,500	09/09/10	20	72	21.88	3.3	1.3%	No	14.3%	YES	NO	4.0×10^{-12}
9	2012	31	8	1	4,752	02/15/12	8	34	9.76	3.5	0.7%	No	0.0%	NA	NO	2.3×10^{-4}
10	2012	207	1	4	1,549	03/23/12	28	31	9.48	3.3	2.0%	Yes[d]	8.5%	YES	YES	5.2×10^{-3}
11	2012	452	1	4	2,411	03/27/12	15	23	6.16	3.7	1.0%	No	5.7%	YES	NO	3.0×10^{-2}
7	2012	53	6	1	1,933	12/03/12	12	44	8.62	5.1	2.3%	No	2.1%	NA	NO	1.5×10^{-12}
8	2012	673	1	4	3,628	12/03/12	14	48	13.25	3.6	1.3%	No	1.2%	NA	NO	2.3×10^{-8}
Total					**30,946**		**177**	**439**								

Clusters presented in the table were selected with the following criteria: cluster duration <7 days, excess cases >10, ratio observed/expected cases of AGI >3, *p*-value <0.05.

NA, Not available.

[a] The medication rate was estimated for the total population of municipalities impacted.

[b] Percentage of analysis >1 *E. coli* and/or fecal streptococci for the period 2009–2012.

[c] 900 *E. coli* UFC/100 mL – 21/06/10.

[d] >100 *E. coli* UFC/100 mL – 10/04/12.

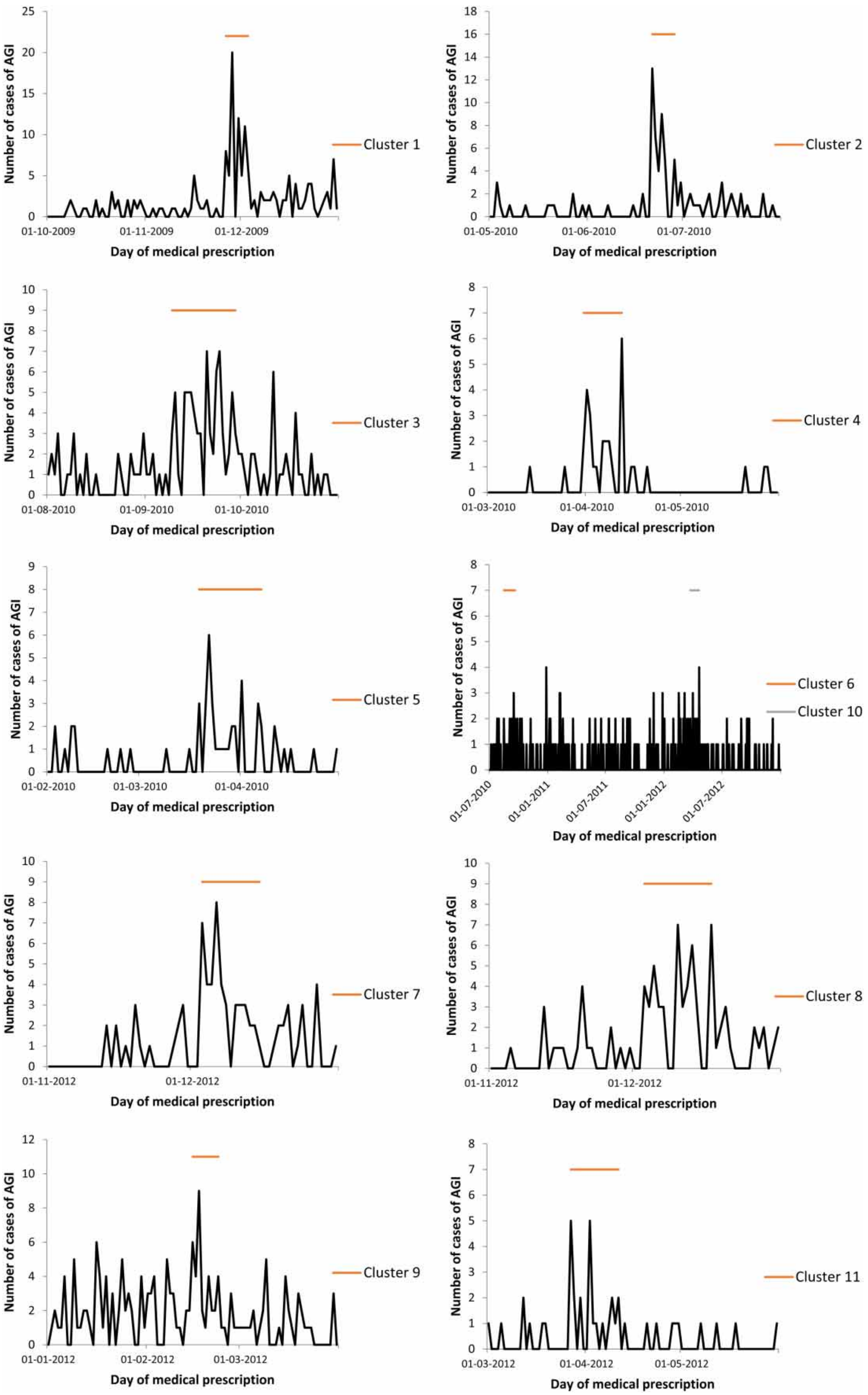


Figure 6 | Description of the number of cases of AGI according to the day of medical prescription and selected clusters.

clusters, fecal pollution of DWN during the outbreak (clusters 2 and 10 in Table 2) was detected. For the same two clusters, a notification of WBDO had been made to the local health authority. Moreover, one geographic area (Group ID 207 in Table 2) concerned two clusters, respectively, in 2010 (cluster 6) and in 2012 (cluster 10). The sanitary control of fecal indicators in drinking water highlighted the occurrence of several episodes of non-microbiological compliance between 2009 and 2012 for nine of the 11 clusters detected (mean = 5.6%; min = 0%; max = 14.3%).

Environmental risk factors of pollution of DWN were identified for all selected clusters when associated information was available (64% of clusters). These included heavy rains in the days before the start of the outbreak (clusters 2, 3, 4, 5, 10, 11), the flooding of the drinking-water borehole causing a cessation of chlorination (clusters 2 and 10), and water pipe breakage in a DWN (cluster 1).

Finally, except for cluster 9, at least one environmental factor (at least percentage of non-microbiological compliance; several missing data have been observed on other information) or water treatment/distribution incident was associated with clusters selected as WBDO.

DISCUSSION

Several factors can explain the occurrence of clusters of AGI cases and their increased incidence. The most commonly documented factors are the ingestion of contaminated food (foodborne disease), person-to-person transmission (in particular in children and older populations), and WBDO. The integrated approach developed in this article for the detection of WBDO using health administrative databases identified 11 AGI clusters in the Auvergne region between 2009 and 2012 where a link with the consumption of contaminated tap water was likely. However, although the integrated approach optimizes the reliability of this link (by taking into account the DWN area prior to detection), all identified clusters have to be analyzed and investigated individually to increase the accuracy of determination.

Validation of selected clusters as WBDO

Several criteria (statistical, epidemiological, and environmental) pointed to the existence of WBDO for the 11 selected clusters:

First, cases of AGI which shared the same DWN and therefore had homogenous drinking water quality were aggregated into newly created geographic areas (by the algorithm) before the application of a space-time detection method. This ecological approach helped highlight any link between health signal (cluster of AGI) and exposure factor (DWN). Nevertheless, as seen in Table 1, 21.8% of the study population lives in municipalities served by more than one DWN (configuration 1 municipality = n DWN). For this configuration, the unit of aggregation of cases of AGI is the municipality. Health data do not enable us to geo-localize cases of AGI at an infra-municipality level. For the seven clusters selected where one municipality was served by more than one DWN, additional investigation is needed to identify the impacted DWN. This would include checking for incidents in water treatment processes and in the distribution networks.

For epidemiological evidence, we used results from past investigations of WBDO (Beaudeau *et al.* 2008) in impacted populations to identify several criteria for selecting clusters as follows: they usually last 1 to 3 weeks (clusters over 6 days were selected here), at least a few dozen cases are involved (clusters with more than ten cases of AGI were selected here), the relative risk presented is greater than 3 (the same value was used here). Finally, a *p*-value <0.05 was also chosen. Moreover, a recent comparative study for the description of two WBDO by using two data sources (cohort study and health administrative database) highlighted a low medication rate in the population (1.5% and 2%, respectively, for both WBDO) (Mouly *et al.* 2016). The medication rate observed for selected clusters in the present study was between 0.7% and 4.8%) (Table 2). Moreover, the application of epidemiological criteria enabled us to exclude other origins of localized outbreak of AGI, for example, foodborne origins, usually characterized by an outbreak duration between 1 and 7 days, and most of the time by fewer than ten cases.

In addition to these epidemiological criteria, we looked for environmental factors for each selected cluster. The occurrence of WBDO is often associated with heavy rainfall (Beaudeau *et al.* 2008), particularly in rural areas with small DWN exposed to fecal discharge from livestock farms. Furthermore, the boreholes of small DWN are often poorly protected compared with DWN in urban areas. For example, in cluster 3, a hydrological report indicated that heavy rains fell on 7 September 2010, i.e., 2 days before the start of a detected cluster. For clusters 2 and 10, two WBDO were investigated and are described in detail in the literature (Mouly *et al.* 2016).

The combination of the integrated approach, which takes into account exposure to DWN before the detection of clusters of AGI and the application of selection criteria of cluster detected based on epidemiological knowledge, enabled us to improve the overall specificity of our detection method. Moreover, the occurrence of several clusters of AGI at different times, focused on the same DWN (e.g., two clusters for area 207 in 2010 and 2012, Table 2), provided strong evidence of a WBDO.

Additional environmental investigation for DWN associated with selected clusters will be necessary to identify the circumstances and the origin of the contamination of tap water.

Benefits and limits of the algorithm in the context of an integrated detection system for WBDO

Increased likelihood of WBDO detection and additional investigations required

Health data were available for municipalities. Drinking water exposure data depended on the individual DWN. We created an algorithm to take into account exposure to drinking water to use in tandem with an existing method for detecting outbreaks of AGI. AGI clusters detected by the combined system have good specificity with respect to the individual water supply. The links between clusters and drinking water must still be confirmed by environmental investigations (rainfall) and the search for possible incidents in drinking water processes and distribution networks on the date of AGI clusters.

As our definition of a DWN assumes homogenous water quality, if pollution is introduced somewhere into the network, it spreads throughout the whole DWN concerned. However, such a hypothesis does not consider the state of individual pipes, differences in flow rates, and stagnant water which occurs when water is not drawn for a long period of time. A potential bias may also occur when the network is contaminated by waste water reflux. All of these situations imply a great deal of heterogeneity in the water quality, depending on the position of the water treatment plant.

The advantage of using the algorithm-based method is clear when several municipalities are served by a single DWN. In our study, all the AGI cases occurring in the same DWN were considered together. The corresponding configurations, i.e., 'm municipalities = 1 DWN' and 'm municipalities = n DWN', concerned 49.7% of the population and 72.9% of the municipalities (Table 1). On the contrary, the creation of the new geographical areas did not help to determine which DWN was involved when a municipality served by several DWN was concerned by an AGI outbreak. In our study, 41.1% of the municipalities in the new geographical areas were associated with two or more DWN. Nevertheless, the exposure-municipalities and exposure-DWN ratios provided information which helped us to focus further investigations on a specific DWN and to confirm the hydric origin. For the configuration '1 municipality = n DWN', in the case of a disease outbreak, it is not possible to identify the DWN responsible, as AGI cases are counted in a municipality which is bigger than the corresponding DWN. Nevertheless, merging the DWN associated with the same municipality (or grouping of municipalities) helps decrease the number of occurrences of the municipalities in the dataset.

Improving power of detection

The algorithm created 714 new geographical areas. This number is much lower than the number of municipalities and DWN (respectively, 1,310 and 1,706), which implies a shorter computation time of spatiotemporal outbreak detections, because fewer geographical units need to be tested. Moreover, the average population size of the new geographical units was much greater (1,891 inhabitants) than for municipalities (1,031 inhabitants) and DWN (791 inhabitants). In turn, this implies improved power of detection of clusters using the algorithm over the standard approach.

Algorithm characteristics

Over the course of developing the algorithm, several methods seemed suitable. First, we considered that maximizing the number of potential AGI cases (to increase the power of detection) was as important as taking into account the corresponding population's exposure to tap water. Thus, we gave the same importance to the exposure-municipalities and exposure-DWN ratios. Second, we chose to minimize the Euclidian distance between combined ratios (1,1) and (exposure-municipalities ratio, exposure-DWN ratio) to select the grouping of municipalities which best matched the specific DWN. These two choices did not greatly influence the results.

The set of new geographic areas constituted a territory whose characteristics (population size, global incidence, and number of AGI cases) were very close to those of the Auvergne region. Accordingly, any repetition of municipalities had an insignificant impact in the incidence evaluation.

Conditions to apply the algorithm

The study area is characterized by a particularly hilly landscape. The relationship between DWN and municipalities is very complicated, and most of the region is rural. Accordingly, one can suppose that the algorithm can be used in other less topographically complex territories as part of an integrated approach for the detection of WBDO.

The health and environmental data used in the algorithm are available for all French regions, so the integrated approach developed here for the detection of WBDO can be applied to other regions in France.

SISE-Eaux database quality

The SISE-Eaux database is maintained at a regional and departmental level. The reliability of this database is essential to obtain accurate matching of drinking water exposure and AGI cases. These data are very reliable in the area studied (Auvergne), in particular the population size counted at the overlap of municipalities and the DWN, an element which is of crucial importance when applying our algorithm.

Space-time detection method and setting

The space-time detection method used to detect clusters of AGI sharing the same DWN (Kulldorff 2010) was selected both because of the consideration of seasonality and the simplicity of its application with SatScan software. With respect to the former, selected clusters after analysis were as numerous during the winter season (5/11 clusters between January and March) as the rest of the year. While a high incidence of AGI is common in European countries during winter, the space-time detection method does not appear to have been influenced by this phenomenon. For time aggregation, we decided to use 'days' whereas most retrospective studies use 'weeks' or 'months' (Demattei 2006). This decision was based on the high incidence of AGI compared with other infectious or chronic diseases. Day-based aggregation time ensures day time precision for detected cluster duration.

Implication for waterborne disease detection

The challenge of WBDO detection addressed in published studies (Edge *et al.* 2004; Berger *et al.* 2006; Andersson *et al.* 2014) highlights the difficulty of detecting short outbreaks involving fewer than 100 cases. For this purpose, information collected for cases has to have sufficient temporal (ideally daily) and spatial (municipality level may be sufficient) resolution to enable the detection of local outbreak signals like WBDO. Unlike other studies, the clusters identified in our study involved fewer than 100 cases of AGI. This would suggest that our method has good sensitivity.

In addition, syndromic surveillance is useful to estimate the size, duration, and health impact of detected outbreaks, as it provides us with the consultation rate in the impacted population. Any such estimation should take into account influencing factors on consultation rate, in particular age and access to health services, as shown in our study, and described elsewhere (Mouly *et al.* 2016).

From a public health point of view, detected epidemic signals from SNIIRAM data should be followed by implementing a set of operational measures, including field investigation. These should be conducted to validate and describe the outbreak, and to understand the origin and mechanisms involved in case diffusion. In turn, this information can inform decision-making for public health prevention.

CONCLUSION

We implemented an algorithm to create new geographical areas which matched health data and environmental exposure levels in drinking water, despite complicated associations between municipalities and DWN. The 714 new geographical areas/units accounted for all the DWN and municipalities in the Auvergne region. The new geographic areas were bigger than the DWN and municipalities, both in terms of surface and population sizes. Creating these areas resulted in greater power of detection of potential future outbreaks. The application of a space-time detection method on the new geographical areas for the Auvergne region between 2009 and 2012 identified 11 potential WBDO.

Accordingly, the relevance of this approach needs to be strengthened by analyzing other datasets (as described in this article).

ACKNOWLEDGEMENTS

The authors thank the National Health Insurance for access to health administrative data. We also thank Pascal Beaudeau, Henriette De Valk and Yann Le Strat from the French National Public Health Agency for their advice and recommendations, and Magali Corso from the French National Public Health Agency for the preparation of case data of acute gastro-enteritis from health administrative databases and Jerôme Pouey for his help in maps conception. Finally, we also thank the Health Ministry and Henri Davezac for water data drawn from the Sise-Eaux database and to Farida Mihoub and Jude Sweeney for her help in translation.

REFERENCES

Andersson, T., Bjelkmar, P., Hulth, A., Lindh, J., Stenmark, S. & Widerstrom, M. 2014 Syndromic surveillance for local outbreak detection and awareness: evaluating outbreak signals of acute gastroenteritis in telephone triage, web-based queries and over-the-counter pharmacy sales. *Epidemiol. Infect.* **142**, 303–313.

Assuncao, R. & Correat, T. 2009 Surveillance to detect emerging space-time clusters. *Comput. Stat. Data Anal.* **53**, 2817–2830.

Balter, S., Weiss, D., Hanson, H., Reddy, V., Das, D. & Heffernan, R. 2005 Three years of emergency department gastrointestinal syndromic surveillance in New York City: what have we found? *MMWR Suppl.* **54**, 175–180.

Beaudeau, P., De Valk, H., Vaillant, V., Mannschott, C., Tillier, C., Mouly, D. & Ledrans, M. 2008 Lessons learned from ten investigations of waterborne gastroenteritis outbreaks, France, 1998–2006. *J. Water Health* **6**, 491–503.

Berger, M., Shiau, R. & Weintraub, J. M. 2006 Review of syndromic surveillance: implications for waterborne disease detection. *J. Epidemiol. Community Health* **60**, 543–550.

Bounoure, F., Beaudeau, P., Mouly, D., Skiba, M. & Lahiani-Skiba, M. 2011 Syndromic surveillance of acute gastroenteritis based on drug consumption. *Epidemiol. Infect.* **139**, 1388–1395.

Chaput, E. K., Meek, J. I. & Heimer, R. 2002 Spatial analysis of human granulocytic ehrlichiosis near Lyme, Connecticut. *Emerg. Infect. Dis.* **8**, 943–948.

Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., Calderon, R. L., Roberts, J. M., Beach, M. J. & Roy, S. L. 2010 Causes of outbreaks associated with drinking water in the United States from 1971 to 2006. *Clin. Microbiol. Rev.* **23**, 507–528.

Cucala, L. 2009 A flexible spatial scan test for case event data. *Comput. Stat. Data Anal.* **53**, 2843–2850.

D'Aignaux, J. H., Cousens, S. N., Delasnerie-Laupretre, N., Brandel, J. P., Salomon, D., Laplanche, J. L., Hauw, J. J. & Alperovitch, A. 2002 Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998. *Int. J. Epidemiol.* **31**, 490–495.

Demattei, C. 2006 *Détection d'agrégats temporels et spatiaux [Space-time clusters detection]*. PhD thesis. Université de Montpellier 1, Montpellier, France.

Edge, V. L., Pollari, F., Lim, G., Aramini, J., Sockett, P., Martin, S. W., Wilson, J. & Ellis, A. 2004 Syndromic surveillance of gastrointestinal illness using pharmacy over-the-counter sales. A retrospective study of waterborne outbreaks in Saskatchewan and Ontario. *Can. J. Public Health* **95**, 446–450.

Fukuda, Y., Umezaki, M., Nakamura, K. & Takano, T. 2005 Variations in societal characteristics of spatial disease clusters: examples of colon, lung and breast cancer in Japan. *Int. J. Health Geogr.* **4**, 16.

Hayran, M. 2004 Analyzing factors associated with cancer occurrence: a geographical systems approach. *Turkish Journal of Cancer.* **34**, 67–70.

Heffernan, R., Mostashari, F., Das, D., Besculides, M., Rodriguez, C., Greenko, J., Steiner-Sichel, L., Balter, S., Karpati, A., Thomas, P., Phillips, M., Ackelsberg, J., Lee, E., Leng, J., Hartman, J., Metzger, K., Rosselli, R. & Weiss, D. 2004 New York City syndromic surveillance systems. *MMWR Morb Mortal Wkly Rep.* **53** (Suppl.), 23–27.

Hrudey, S. E. & Hrudey, E. J. 2004 *Safe Drinking Water: Lessons from Recent Outbreaks in Affluent Nations*. IWA Publishing, London, 486 pp.

Institut National de l'Information Géographique et Forestière 2013 Available from http://professionnels.ign.fr/geofla.

Klassen, A. C., Kulldorff, M. & Curriero, F. 2005 Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. *Int. J. Health Geogr.* **4**, 1.

Kulldorff, M. 2010 StaScan User Guide for version 9.0. 110.

Kulldorff, M., Feuer, E. J., Miller, B. A. & Freedman, L. S. 1997 Breast cancer clusters in the northeast United States: a geographic analysis. *Am. J. Epidemiol.* **146**, 161–170.

Kulldorff, M., Heffernan, R., Hartman, J., Assuncao, R. & Mostashari, F. 2005 A space-time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2**, e59.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R. & Kulasekera, V. 2003 Dead bird clusters as an early warning system for West Nile virus activity. *Emerg. Infect. Dis.* **9**, 641–646.

Mouly, D., Van Cauteren, D., Vincent, N., Vaissiere, E., Beaudeau, P., Ducrot, C. & Gallay, A. 2016 Description of two waterborne disease outbreaks in France: a comparative study with data from cohort studies and from health administrative databases. *Epidemiol. Infect.* **144**, 591–601.

Odoi, A., Martin, S. W., Michel, P., Middleton, D., Holt, J. & Wilson, J. 2004 Investigation of clusters of giardiasis using GIS and a spatial scan statistic. *Int. J. Health Geogr.* **3**, 11.

Osei, F. B. & Duker, A. A. 2008 Spatial dependency of V. cholera prevalence on open space refuse dumps in Kumasi, Ghana: a spatial statistical modelling. *Int. J. Health Geogr.* **7**, 62.

Patil, G. P. & Taillie, C. 2004 Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environ. Ecol. Stat.* **11**, 183–197.

Takahashi, K., Kulldorff, M., Tango, T. & Yih, K. 2008 A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *Int. J. Health Geogr.* **7**, 14.

Tuppin, P., De Roquefeuil, L., Weill, A., Ricordeau, P. & Merliere, Y. 2010 French national health insurance information system and the permanent beneficiaries sample. *Rev. Epidemiol. Sante Publique* **58**, 286–290.

First received 2 November 2015; accepted in revised form 6 February 2017. Available online 25 March 2017