An extreme value theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of Salmonella

Armelle Guillou, Marie Kratz, Yann Le Strat

An Extreme Value Theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of *Salmonella*

A. Guillou *, M. Kratz † , Y. Le Strat ‡

Abstract

We propose a new method which could be part of a warning system for the early detection of time clusters applied to public health surveillance data. This method is based on the extreme value theory (EVT). To any new count of a particular infection reported to a surveillance system, we associate a return period which corresponds to the time that we expect to be able to see again such a level. If such a level is reached, an alarm is generated. Although standard EVT is only defined in the context of continuous observations, our approach allows to handle the case of discrete observations occurring in the public health surveillance framework. Moreover it applies without any assumption on the underlying unknown distribution function. The performance of our method is assessed on an extensive simulation study and is illustrated on real data from Salmonella surveillance in France.

*Univ. de Strasbourg & CNRS, IRMA UMR 7501, France; Email: armelle.guillou@math.unistra.fr

†ESSEC Business School, CREAR risk research center, France; Email: kratz@essec.edu. M.Kratz is also member of MAP5, UMR 8145, Univ. Paris Descartes, France.

‡French Institute for Public Health Surveillance, Saint-Maurice, France; Email: y.lestrat@invs.sante.fr

1 Introduction

Since the pioneering work of Serfling ([1]), several statistical models have been proposed to detect time clusters from surveillance data recorded as time series count in a given geographic area. A time cluster is defined as a time interval in which the observed number of events is significantly higher than the expected number of events. The term 'event' is generic enough to include any event of interest such as a case of illness, an admission to an emergency department, a death or any other health event.

The early prospective detection of time clusters represents a statistical challenge as the models must take the main features of the data into account such as secular trends, seasonality, past outbreaks but are also faced with idiosyncrasies in reporting, such as delays, incomplete or inaccurate reporting or other artefacts of the surveillance systems. Reporting delays are particularly problematic for surveillance systems that are not based on electronic reporting. Concerning syndromic surveillance systems, the same difficulties are encountered, excepted for the reporting delays because these surveillance systems are mostly based on electronic reporting.

The main statistical approaches introduced in the literature have been recently reviewed by Unkel *et al* [2] with a discussion of statistical issues involved in evaluating and comparing methods. As mentioned by Unkel *et al* [2], the published models are sometimes presented into broad classes such as regression techniques, time series methodology or statistical process control ([3]; [4]; [5]) but this presentation is somewhat artificial as several methods can be classified into more than one of these classes. However, in most cases they are based on two steps: (i) the calculation of an expected

[0]This work was partially supported by the ESSEC Research Center, project #043.203.5.1.0.7.07.

2000 AMSC Primary: 62G32 ; Secondary: 62G05, 62G30
Keywords: Extreme value theory; Return level; Return period; Outbreak detection; Salmonella; Surveillance.

value of the event of interest for the current time unit (generally a week or a day); (ii) a statistical comparison between this expected value and the observed value. A statistical alarm is triggered if the observed value is noticeably different from the expected value.

The first step is based on the past counts, or more often on a sample of the past counts, that takes the seasonality pattern(s) into account. Thus, the current count is compared to counts that occurred in the past during the same time periods, e.g. 3 weeks around the current week from 5 previous years. Alternatively, sinusoidal seasonal components can be incorporated into regression models to deal with the seasonality and to easily take secular trends into account. More rarely, models try to reduce the influence of weeks coinciding with past outbreaks. One solution to avoid that such outbreaks reduce the sensitivity of the model is to associate low weights to these weeks in the estimation of the expectation (e.g. [6]).

The intentional release of anthrax in the USA in October 2001 emphasized the need to develop new early warning surveillance systems ([7]; [8]). These surveillance systems treat an increasing number of data provided from multiple sources of information ([9]). One logical consequence is to perform statistical analyses with a daily frequency.

Developing automated statistical prospective methods for the early detection of time clusters is thus essential. It is important for a public health surveillance agency to run several statistical methods concomitantly in order to compare the alarms generated by these methods.

To this aim, we propose in this paper a new approach based on Extreme Value Theory (EVT) (e.g. [10]) for the early detection of time clusters. EVT has not been used so far in the context of surveillance; adapting and combining in a specific way EVT tools introduces an innovative and interesting alternative to existing methods for surveillance data, not so numerous. One of the main advantages of the EVT approach is that we do not need any model to fit the time series and no assumption is required on the

underlying distribution.

The main idea of our method is the following. First, we associate to each observation of the dataset a return period, which corresponds to the time that we expect to be able to see again such a level (and not earlier). Then, whenever we have a new observation, we look backward on its return period to see if already at least one observation exceeds it. If it does, we consider the new observation as "abnormal" (e.g. the potential start of an epidemic), in the sense of "not expected" as defined by our rule, and we set off an alarm.

To evaluate the method, we ran it on simulated data generated by an extensive simulation study based on recent work [11]. Then to illustrate the method, we applied it to the detection of time clusters from weekly counts of *Salmonella* isolates reported to the national surveillance system in France. Salmonellosis is a major cause of bacterial enteric illness in both humans and animals, with bacteria called *Salmonella*. In France, *Salmonella* is the first cause of laboratory confirmed bacterial gastroenteritis, of hospitalization and of death. In 2005, a study estimated that between 92 and 535 deaths attributable to non typhoidal *Salmonella* occurred each year ([12]).

The paper is organized as follows. The surveillance system and the data are presented in Section 2. A description of our method to check if each new observation corresponds to an unusual/extremal event is given in Section 3. Applications to simulated data and real counts of *Salmonella* are developed in Section 4. A discussion follows in the last section.

2 Data

We will consider, in our paper, simulated datasets in order to test the performance of our method and real datasets in order to illustrate it. Concerning the latter, let us present in this section the real data, the counts of *Salmonella*, on which we will be working. The National Reference Center for *Salmonella* contributes to the surveillance

of salmonellosis by performing serotyping of about 10000 clinical isolates each year. *Salmonella* surveillance is based on a network of 1500 medical laboratories that voluntarily send their isolates. *Salmonella* enterica serotypes Typhimurium and Enteritidis represent 70 % of all *Salmonella* isolates in humans among many hundreds of serotypes; that is why we consider in this paper mainly these two serotypes. For illustrative purpose, four other less frequent serotypes (Manhattan, Derby, Agona and Virchow) are also considered; Figure 1 shows the weekly number of isolates for these six serotypes from January 1, 1995 to December 31, 2008. It highlights the great variability in terms of seasonality, secular trend and weekly number of reported isolates and frequencies of unusual events. The number of cases is recorded by week of reporting to the National Reference Center because some dates of typing are missing.

Figure 1: Weekly counts of isolates reported to the National Reference Center for *Salmonella* in France, January 1, 1995 to December 31, 2008: (a) *Salmonella* Manhattan; (b) *Salmonella* Derby; (c) *Salmonella* Agona; (d) *Salmonella* Virchow; (e) *Salmonella* Typhimurium; (f) *Salmonella* Enteritidis.

Let $Y = \{Y_t; t = 1, \ldots, T\}$ be the univariate initial time series corresponding to the number of isolates at time point t for a given serotype. If the series Y exhibits a trend, we estimate it from a linear regression trend component, then we substract it.
Concerning the possible seasonality of Y, as mentioned by many authors (e.g. [7]; [13]), seasonal effects may have a strong impact in generating a statistical alarm. A common way to prepare the dataset is to select counts from comparable periods in past years, as described in the literature ([6]; [14]). The dataset is restricted to the counts that occurred during the times within these comparable periods. For instance, if the current time is t of year y, then only the counts for the $n = b(2w+1)$ times from $t-w$ to $t+w$ of years from $y-b$ to $y-1$ ($b > 1$, $w > 1$) are used. Another way to handle seasonality is to include seasonal factors in the statistical model running on a restricted dataset, as recently proposed by Noufaily *et al.* [11].

As an illustration Figure 2 represents the restricted dataset for *Salmonella* Typhimurium, for a given current week with $w = 3$ and $b = 5$.

Figure 2: Weekly counts of isolates for *Salmonella* Typhimurium. The current week, represented by the arrow, is the last week of December 2008. The counts that occurred in comparable periods ($\pm$ 3 weeks) in the five previous years, and used to generate or not an alarm, are represented by the striped bands.

In all the sequel, $X = (X_t)$ will denote the transformed (i.e. the detrended subset) time series from (Y_t). To each new observation will correspond a new sample, indifferently named X or (X_t), to simplify. Each step of the method will apply on X itself.

3 An EVT approach

Suppose we have at our disposal n successive observations that we consider as realizations of a sample (X_i) of independent and identically distributed (i.i.d.) non-negative random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, from an unknown distribution function F. We do not make any assumption on F and do not need any to develop our method, which constitutes a great advantage of this approach.

We want to detect extreme events, which justifies an EVT framework. Extreme events have been extensively studied in the literature (e.g. [10]), with two main approaches developed in the i.i.d. setting, one based on the distributions of maxima (e.g. [15]) and the other one based on exceedances above some threshold, known as the Peak-Over-Threshold (POT) method (e.g. [16]).

Assuming independence might appear as a rough approximation to analyze epidemiological discrete time series; however we obtain quite reasonable and interesting results under this assumption, usual when introducing an EVT approach. Note also that in most of the existing methods for surveillance data the residuals from the models are assumed to be independent.

Moreover, the method we propose also allows handling the case of a discrete phenomenon, whereas the classic EVT method requires continuous distributions only.

We will face this issue of using extreme tools for discrete data in one step of our method, when having to estimate extreme quantiles (or return levels); instead of estimating them explicitly as we do in the continuous framework, we will estimate their upper bounds, using bounds introduced in [17] which do not require any strong assumption (such as continuity) on the tail behavior of the distribution function. The practitioner, e.g. a biostatistician or an epidemiologist, is often ready to accept more alarms if he knows that the upper bound is well estimated and can provide a reasonable 'worst case scenario'. In that sense, our method, in particular the use of an upper bound for the return level, should be viewed as an alternative tool which provides additional information for the practitioner.

Now let us present the fundamentals of our method.

Associated with a given return period T which corresponds to T time units over the past, a return level z_T is defined as the level expected to be exceeded on average once every T time units, *i.e.* such that

$$\mathbb{E}\left(\sum_{i=1}^{T} \mathbb{1}_{\{X_i > z_T\}}\right) = 1 \tag{1}$$

where $\mathbb{1}_{\{A\}}$ represents the indicator function that is equal to 1 if A is true and to 0 otherwise. The last equality can be rewritten as $1 - F(z_T) = 1/T$. Hence, the return level z_T corresponds simply to a p_T-quantile with $p_T = 1 - 1/T$, $z_T = F^{\leftarrow}(1 - 1/T)$, $F^{\leftarrow}$ denoting the generalized inverse function of F.

Notice that in (1), we may have chosen to replace the right-hand side of the equation by any small integer, say c. The value of c is related to the height of the threshold above which observations are considered as exceedances (above this threshold). The smaller c, the higher the threshold. If we want to select what we call extreme/unusual values,

it is preferable to select a high threshold, which means a small value of c. We chose here to make it really extreme, not asking for more than one exceedance, in average, above the last observation. We could relax a bit this choice, by taking for instance $c = 2$, but we prefer to be on the safe side. As mentioned in the discussion of Section 5, an optimal value of c, or a prescribed value in order to match the performance of other methods already proposed in the literature is an interesting open question which will lead to further investigations.

The idea of the method is to associate with each observation x_s a return period T_s defined theoretically as $(1 - F(x_s))^{-1}$ to be able to determine the return period T_t associated to each new observation x_t at time t. Then, we look backwards (and not forwards as in the 'standard' way) in the interval $(t - T_t; t)$ for the existence of an observation that would exceed x_t. If it exists, we set the rule to generate an alarm at time t based on the fact that, on average, we do not expect two or more exceedances on $(t - T_t; t]$.

Notice that in our discrete framework it will not be possible to estimate explicitly the return levels; instead estimated bounds will be considered.

Therefore, after a preliminary analysis of the data and definition of our sample, we will compute the estimated bounds of the return levels in order to obtain a graph of the return periods and levels. Then, we will allocate a return period to any new observation x_t to test if t corresponds to a warning time according to our definition.

Looking at extremal events leads us to the crucial problem of high quantile estimation, well-studied in the EVT literature (e.g. [10]), based on a sample of observations. However, EVT is only valid in the case where the underlying distribution function F of these observations is continuous. This is not the case in the surveillance context where the observations are counts and thus F is discrete. Therefore, we propose to use instead upper and lower bounds for the return level z_t and estimate them as in [17]; this approach has several advantages: the upper and lower bounds can be computed for

any value of t (in particular it holds for large values), it does work for both small and large samples, and for F continuous or discrete. So it is well-adapted to our context, when assuming the random variables associated to the i.i.d. observations.
Note that using bounds for a return level z_t will imply that the return period defined theoretically by $(1 - F(z_t))^{-1}$ cannot be explicitly estimated and we have

$$T_\ell \leq (1 - F(z_t))^{-1} \leq T_b \tag{2}$$

where T_ℓ and T_b denote the return periods of the lower and upper bounds $\ell_t(u, w, q)$ and $b_t(u, v)$ respectively, defined in the on-line Web Appendix and estimated below. Here u, v and w are suitable power functions and $q > 1$.

Now let us present our method to define an alarm system. For each new observation x_{t_i}, $i \geq 1$:
• First we check if the new observation is the largest on all the sample. If it is the case, we generate an alarm. This conservative condition may be weakened, applying it only when observing a local non decreasing trend.
• Otherwise, we proceed as follows.
Step 1: We draw the plot of the return period on the x-axis and the corresponding estimate of the upper bound of the return level (instead of the return level itself): $\left(t, \widehat{b}_t\right)$, where

$$\hat{b}_t = \hat{b}_t\left(\widehat{\alpha}_t, \widehat{\beta}_t\right) = \left(\frac{t\hat{\theta}_n\left(\widehat{\alpha}_t, \widehat{\beta}_t\right)}{(1 - 1/t)^{\widehat{\beta}_t}}\right)^{1/\widehat{\alpha}_t}, \quad \text{with} \quad \hat{\theta}_n(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n} x_{i,n}^{\alpha}(i/n)^{\beta}. \tag{3}$$

Step 2: We allocate to each observation x_{t_i}, $i \geq 1$, a return time T_i using the previous plot. Namely, x_{t_i} corresponds to a value $\widehat{b}_{T_i}$ of the y-axis of the plot from which we

deduce the associated return level T_i. Reading an observation as an upper bound of a return level means that T_i is in fact the lower bound of the theoretical return period $(1 - F(x_{t_i}))^{-1}$ that should be associated to the observation x_{t_i}, because of (2).

We justify our choice as follows. Considering $\widehat{\ell}_{T_i}$ instead of $\widehat{b}_{T_i}$ in the above method would have led to underestimate the return period associated to the observation x_{t_i}, which could be a problem in the context of alarms (it is better to have more alarms than less), except if the plots $\left(t, \widehat{\ell}_t\right)$ and $\left(t, \widehat{b}_t\right)$ were close enough, but it is generally not the case for our datasets where $\widehat{\ell}_t$ appeared approximately as a constant function of time ([18]).

Step 3: We use the fact that if (X_j) are i.i.d. random variables, then we have for any time interval $I(T)$ with length T

$$\mathbb{E}\left(\sum_{i=1}^{T} \mathbb{1}_{\{X_i > z_T\}}\right) = 1 \Leftrightarrow \mathbb{E}\left(\sum_{i \in I(T)} \mathbb{1}_{\{X_i > z_T\}}\right) = 1. \tag{4}$$

This remark is important since we want to define for each new observation a warning time, which means to look backward in time.

Hence for each new observation x_{t_i}, to which a return time T_i has been associated (via Step 2), we will look in the interval $(t_i - T_i; t_i)$ to see if there exists an observation exceeding x_{t_i}, considered as the new exceedances threshold; if it does, we ring an alarm at this time t_i.

• To finish this section, let us summarize our method as an iterative algorithm.

For each new observation x_{t_i}, $i \geq 1$:

1. If $x_{t_i} > \max_{j<t_i} x_j$, then generate an alarm time at time t_i. Note that this conservative condition may be weakened, applying it only when observing a local non decreasing trend.

2. Otherwise,

(a) associate the time T_i to x_{t_i}, read from the plot $\left(t, \widehat{b}_t\right)$ when x_{t_i} is regarded as a return level $\widehat{b}_{T_i}$;

(b) consider $I(T_i) = (t_i - T_i, t_i]$;

(c) look for the existence of an observation $x_t \geq \widehat{b}_{T_i}$, for $t \in (t_i - T_i, t_i)$;
if there exists at least such an observation, then generate a warning time at t_i.

Remarks: in practice, if the return period of the new observation x_{t_i} is larger than the length of the sample (X_t), then we choose to be conservative by generating an alarm at time t_i. Other choices may of course be considered. Also, in some sense, detecting outbreaks when two or more "extreme" observations in a period are observed, instead of one, might be considered as a "heuristic rule". But, as illustrated in the next section, our detections are in accordance with the ones obtained with other classical approaches already used in the literature. Nevertheless, a possible alternative, which would require further investigations, might consist in replacing the expectation by an estimation of the probability $\mathbb{P}\left(\sum_{j \in I(T_i)} \mathbb{1}_{\{x_j \geq \hat{b}_{T_i}\}} > 1\right)$.

• Now to illustrate our method, let us consider the example of the number of *Salmonella* Virchow isolates. In Figure 3, the x-axis corresponds to the values of t from 2 to 100 weeks and the y-axis to $\widehat{b}_t$, calculated for the last week of 2008; the two dashed lines indicate the 95% confidence interval bounds of b_t.

Figure 3: The return level/return period graph for *Salmonella* Virchow, calculated for the last week of 2008. The black curve represents the upper bound of the return level. Dashed curves represent the 95% confidence interval of this upper bound. To the observation $y = 20$ (respectively $y = 15$) does correspond on the x-axis $\widehat{b}_{91}$ (respectively $\widehat{b}_7$) from which we deduce the return period equals to $T = 91$ (respectively $T = 7$) weeks.

4 Applications

4.1 Simulated data

We generated simulated data following the approach proposed by Noufaily *et al.* [11]. Let us summarize briefly the procedure:

- First, simulated baseline data (i.e. time series of counts in the absence of outbreaks) were generated using a negative binomial model with a mean including trend and seasonality determined by Fourier terms. More precisely, these data were simulated from 42 different parameter combinations (called scenarios and presented in Table 1 in [11]) with different trends, seasonalities, baseline frequencies of counts and dispersions. The simulations of the baseline data use 100 replicates from each scenario of size 624 weeks. The last 49 weeks were used as current weeks leading to 4900 replicates for each of the 42 scenarios.

- Secondly, outbreaks were simulated both in baseline and current weeks. Given a constant k, the size of a simulated outbreak, starting in week t_i, follows a Poisson distribution with mean equal to k times the standard deviation of the count at t_i. Then outbreak cases were randomly distributed according to a lognormal distribution with mean 0 and standard deviation 0.5. Finally, four outbreaks were generated in baseline weeks and one outbreak was generated in current weeks. Outbreaks start times were chosen randomly. We chose the values of k to be 3 and 5 in baseline weeks and from 1 to 10 in current weeks.

- Next, we used two measures, named FPR and POD, introduced in [11] to assess the performance of the EVT method. For each scenario, we calculated the false positive rate (FPR) as the proportion of the current 49 weeks and 100 replicates, in which the method generated an alarm in the absence of outbreak. The second measure, called the probability that an outbreak is detected (POD) was also

calculated. For each scenario and for each current week, if an alarm is generated at least once between the start and end times of the outbreak, the outbreak is considered as detected. The POD is the proportion of outbreaks detected in 100 runs.

Figure 4: False positive rates obtained for outbreaks of 3 (a) standard deviations and of 5 (b) standard deviations. Proportions of detected outbreaks of k standard deviations corresponding to the 42 scenarios. Outbreaks of 3 (c) and 5 (d) standard deviations are included in baselines.

Results of this simulation study are presented in Figure 4 showing the FPRs and the POD when outbreaks of 3 (plots (a), (c)) and 5 (plots (b), (d)) standard deviations are included in baselines. We show the central estimates of the FPRs associated to their 95% confidence intervals. As noted in [11], the FPRs are highest for scenarios 10-12 with low baseline and over-dispersed data. In our study, the FPR is also high for scenario 16. Concerning the POD, as expected it increased with k for each scenario. The PODs are not so different when outbreaks of 3 standard deviations and 5 standard deviations are present in the baselines.

- Finally, we compared the results obtained for each of the 42 simulated scenarios respectively with the EVT approach and with the Model 0 of Farrington's method (see Noufaily et al., 2012) run with $\alpha = 1\%$. We chose the Farrington method for two main reasons. First the aim of the Farrington algorithm was to develop a robust method for the routine monitoring of weekly reports on infections for many different pathogens at the Communicable Disease Surveillance Center in UK. This method was applied in particular to the detection of Salmonella outbreaks. Second, this method has been applied in France for the surveillance of human Salmonella since 2006 and for the surveillance of Salmonella isolated in the agro-food chain in France for the last two years ([19]) and has shown good

performances. In 2008, performances of five methods were evaluated for the early detection of excess legionella cases in France ([20]) concluding that the Farrington method had the best positive predictive value, equal to 67%. More specifically, we presented in Figure 5 the POD obtained from the 100 runs for each scenario with the Farrington method when outbreaks of 3 standard deviations are included in baselines. This plot can be compared to Figure 4(c) obtained with the EVT approach. As expected, this proportion increases with k for each scenario. For completeness, we also represented on a same plot the false positive rates obtained respectively with the EVT approach (black circles) and with the Farrington method (white triangles). The means represented by the dots together with the 95% confidence intervals are based again on these 100 replicates. It is fairly clear that the performances of the methods depend both on the characteristics of the time series summarized by the different scenarios and the magnitude of past and current epidemics. However, both methods favor a good specificity rather than high sensitivity, and they have a close performance, with a slightly better one for the EVT.

Figure 5: POD obtained with the Farrington method, Model 0 with $\alpha = 1\%$. Proportions of detected outbreaks of k standard deviations corresponding to the 42 scenarios. Outbreaks of 3 standard deviations are included in baselines.

Figure 6: False positive rates obtained for outbreaks with the EVT approach (black circles) and with the Farrington method, Model 0 with $\alpha = 1\%$, (white triangles), for 42 scenarios.

4.2 Real data

For each week from January 2000 to December 2008, the EVT method was applied to two time series presented in Section 2 (Agona and Manhattan). For each time series Y and for a week t:

1. We note x_t the observed number of cases at week t, that we will compare to selected counts of weeks from $t-3$ to $t+3$ of years from $y-5$ to $y-1$ (to handle the seasonality as explained in Section 2). This restricted dataset corresponds to the resulting time series X on which we apply our EVT method.

2. We calculate the upper bound for x_t regarded as the return level, using equation (3).

3. We calculate a return period associated to the upper bound of the return level. In practice we calculate for $t = 2$ to T_{max} (e.g. $T_{max} = 100$) the increasing upper bounds $\hat{b}_{t=2}, \hat{b}_{t=3}, \ldots, \hat{b}_{t=T_{max}}$. The return period is determined by the rank of the first upper bound greater than the observed number of cases x_t.

4. We generate an alarm at time t if:

   (a) the return period is larger than the length of the sample X, because we have not enough past data to claim that x_t is not an unusual event;

   (b) x_t is larger than the maximum number of counts observed in the sample X;

   (c) we observe, over the return period, a number of cases greater than x_t.

Moreover, in order to reduce the probability that an alarm could be triggered for few sporadic cases, a standard rule, put as an option in the program because specific to Salmonella, has been adopted to keep an alarm at week t if at least 5 cases were observed during the 4 last weeks preceding t. This rule has already been applied at the former Communicable Disease Surveillance Center (CDSC, now Center for Infections

of the Health Protection Agency) in the UK using the method developed by Farrington *et al.* ([6]). It is an empirical rule, which avoids too numerous alarms whenever the time series concerns a rare serotype only. In this specific case, an alternative could be to transform equation (1) into $\mathbb{E}\left(\sum_{i=1}^{T} \mathbb{1}_{\{X_i > z_T\}}\right) = c$ with a suited small value $c > 1$ as previously discussed.

We applied the EVT method on the weekly number of isolates for serotypes Manhattan and Agona. We compared the alarms generated by the EVT method with those generated by the Farrington method (corresponding to *Model 0* in [11] with the threshold value at time t defined as the upper 99% prediction limit).

Figures 7 and 8 represent both the alarms and the weekly counts over time for the serotypes Manhattan and Agona. Each triangle represents a statistical alarm. Triangles on the first line represent the alarms generated by the EVT method, whereas the ones generated by the Farrington method are given on the second line.

Figure 7: *Salmonella* Manhattan: Weekly counts from January 1, 2000 to December 31, 2008. Roman numerals refer to the quarters of the years. Alarms generated by the EVT method are represented by triangles on the first line. Alarms generated by the Farrington method are represented by triangles on the second line. The documented outbreak is delimited by the two dashed lines.

In Figure 7, the alarms generated by the two methods occurred in the same period that corresponds to a documented outbreak, delimited by the dashed lines, for the serotype Manhattan ([21]). From August 2005 to February 2006, a community-wide outbreak of *Salmonella* Manhattan infections occurred in France. The investigation incriminated pork products from a slaughterhouse as being the most likely source of this outbreak. There was a concordance between the temporal (October-December

2005) and the geographical (south-eastern France) occurrence of the majority of cases and the distribution of products from the slaughterhouse.

Figure 8: *Salmonella* Agona: Weekly counts from January 1, 2000 to December 31, 2008. Roman numerals refer to the quarters of the years. Alarms generated by the EVT method are represented by triangles on the first line. Alarms generated by the Farrington method are represented by triangles on the second line. The documented outbreak is delimited by the two dashed lines.

In Figure 8, alarms for the serotype Agona are distributed from 2000 to 2008. There is a concordance between the two methods during three periods. The first period, corresponding to 5 weeks in August and September 2003, was not documented as an outbreak. The second concordance period, corresponds to 15 consecutive weeks from the last week in December 2004 to week 15 in April 2005. This second period is more interesting as it corresponds to the beginning of a large outbreak of infections in infants linked to the consumption of powdered infant formula ([22]). The outbreak period, delimited by the two dashed lines, took place in two stages: the first stage from week 53 in December 2004 to week 10 in March 2005 and the second from week 11 to week 21. A total of 47 cases less than 12 months age were identified during the first stage and 94 cases less than 12 months age were identified during the second stage. The third period corresponds to the week 29 in July 2008. It included five cases, two of them coming from a foodborne disease outbreak involving piglet consumption, and the three others being probably sporadic cases.

The EVT method was implemented using R version 2.9 ([23]). The R-code is available on request. The function called *algo.farrington*, implemented into the R-Package *surveillance* ([24]) was used to apply the Farrington method.

5 Discussion

We believe that the EVT method meets a number of requirements, listed by Farrington *et al.* ([3]), for the outbreak detection algorithms implemented in surveillance systems. Indeed, this method is able to monitor a large number of time series which became an absolute necessity in modern computerized surveillance systems. It can deal with a wide range of events as it is the case for the *Salmonella* infections with the routinely analyses of several hundred serotypes per week. It can handle time series with great numbers of cases (such as *Salmonella* Enteritidis) or small numbers of cases (such as *Salmonella* Manhattan). Seasonality is taken into account by comparing counts over the same periods of time. Other methods propose a direct way to treat the past aberrations, for instance by associating low weights to the weeks coinciding with past outbreaks. There is no such need when using the EVT method since the return period is not a constant but depends on each observed count; alarms can then be generated even if past outbreaks exist. If a past outbreak is contained in the return period interval, then an alarm will be triggered at time t if the observation at t has been exceeded by this outbreak. Finally, the method is implemented in a function using the R language, allowing to run it in an automated procedure with minimal user intervention. This method could be easily included into the R surveillance package and used by public health surveillance practitioners. Recently, several papers have shown that R is routinely used for the early detection of outbreaks in Europe ([25]; [19]; [26]; [27]; [28]).

Although the model developed by Farrington *et al.* ([6]) became a standard reference method, routinely used in France since many years and incorporated in several surveillance systems, the EVT method seems also to be a valuable and interesting tool for the recognition of time clusters. It could be integrated in the family of outbreak detection algorithms used by the public health surveillance agencies since developing effective computer-assisted outbreak detection systems still remains a necessity to en-

sure timely public health intervention.

Obviously our approach does not pretend answering all the issues and could be improved. This paper must be viewed as an exploratory study and constitutes a first step for further developments already quite promising. The main two advantages of our approach is that it does not require any assumption on the underlying (unknown) distribution function and leads to a good specificity. Unfortunately, as the Farrington one, it is less sensitive than other methods. That is why it should not be used solely but combined with other more sensitive methods. Due to its competitiveness against standard approaches, it should help practitioners in the detection of epidemics.

The overall potential of our approach, theoretical as practical, together with a fuller comparison to other methods will be the subject of a forthcoming paper. In particular, an interesting open question is the choice of an optimal value c for the right-hand side in equation (1). This question comes back to discuss the choice of the threshold. In this paper we fixed $c = 1$. As emphasized in the simulation study, this value leads to competitive results of our approach with respect to the Model 0 of Farrington's method in a large panel of situations, in terms of parameter combinations, trends, seasonalities, baseline frequencies of counts and dispersions, but this choice might be improved by an optimal selection. This leads to further investigations.

Acknowledgements

The authors are grateful to an Associate Editor and two anonymous reviewers for their comments and suggestions that led to a marked improvement of the article. They would like to thank Angela Noufaily and Paddy Farrington for providing them simulated datasets and a R code used in Section 4.1 and Francois-Xavier Weill for providing us the *Salmonella* datasets. They also thank Anis Borchani for the implementation of the

method using the R language.

References

[1] Serfling RE. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Reports* 1996; **78**: 494–506.

[2] Unkel S, Farrington P, Garthwaite P, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2012; **175**(1): 49–82.

[3] Farrington CP, Andrews NJ. Statistical aspects of detecting infectious disease outbreaks. In: Brookmeyer, R. and Stroup, D.F. (editors). *Monitoring the Health of Populations* Oxford University Press, 2004; 203–231.

[4] Le Strat Y. Overview of temporal surveillance. In: Lawson, A.B. and Kleinman, K. (editors). *Spatial and Syndromic Surveillance* Wiley, 2005; 13–29.

[5] Sonesson C, Bock D. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A* 2003; **166**: 5–21.

[6] Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society Series A* 1996; **159**: 547–563.

[7] Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. *Proceedings of the National Academy of Sciences of the United States of America* 2002; **99**: 5237–5240.

[8] Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences of the United States of America* 2003; **100**: 1961–1965.

[9] Centers for Disease Control and Prevention. Syndromic Surveillance: Reports from a National Conference 2003. *Morbidity and Mortality Weekly Report* 2004; **53**(Suppl).

[10] Embrechts P, Klüppelberg C, Mikosch T. *Modelling extremal events for Insurance and Finance* 2001; Springer.

[11] Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine* 2012; (online; DOI: 10.1002/sim.5595).

[12] Vaillant V, de Valk H, Baron E, Ancelle T, Colin P, Delmas MC, Dufour B, Pouillot R, Le Strat Y, Weinbreck P, Jougla E, Desenclos JC. Burden of foodborne infections in France. *Foodborne Pathogens and Disease* 2005; **2**: 221–232.

[13] Nobre FF, Monteiro ABS, Telles PR, Williamson GD. Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine* 2001; **20**: 3051–3069.

[14] Stroup DF, Williamson GD, Herndon JL. Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statistics in Medicine* 1989; **8**: 323–329.

[15] Fisher R, Tippett L. Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proceedings Cambridge Philosophy Society* 1928; **24**: 180–190.

[16] Pickands J. Statistical inference using extreme-order statistics. *Annals of Statistics* 1975 **3**: 119–131.

[17] Guillou A, Naveau P, Diebolt J, Ribereau P. Return level bounds for discrete and continuous random variables. *Test* 2009; **18**: 584–604.

[18] Borchani A. Statistiques des valeurs extrêmes dans le cas de lois discrètes. *Rapport ESSAI (2008) & ESSEC Working Paper 10009 (2010).*

[19] Danan C, Baroukh T, Moury F, Da Silva-Jourdan N, Brisabois A, Le Strat Y. Automated early warning system for the surveillance of Salmonella isolated in the agro-food chain in France. *Epidemiology and Infection* 2011; **139**(5): 736–741.

[20] Grandesso F. Early detection of excess legionella cases in France: evaluation performance of five automated methods. *ESCAIDE : European Scientific Conference on Applied Infectious Diseases Epidemiology* 2009.

[21] Noel H, Dominguez M, Weill FX, Brisabois A, Duchazeaubeneix C, Kerouanton A, Delmas G, Pihier N, Couturier E. Outbreak of Salmonella enterica serotype Manhattan infection associated with meat products, France, 2005. *Eurosurveillance* 2006; **11**: 270–273.

[22] Brouard C, Espie E, Weill FX, Kerouanton A, Brisabois A, Forgue AM, Vaillant V, de Valk H. Two consecutive large outbreaks of Salmonella enterica serotype Agona infections in infants linked to the consumption of powdered infant formula. *The Pediatric Infectious Disease Journal* 2007; **26**: 148–152.

[23] R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org. 2006.

[24] Höhle M. Surveillance: An R package for the surveillance of infectious diseases. *Computational Statistics* 2007; **22**: 571–582.

[25] Cakici B, Hebing K, Grünewald M, Saretok P, Hulth A. CASE: a framework for computer supported outbreak detection. *BMC Medical Informatics and Decision Making* 2010; **10**:14 (online;doi: 10.1186/1472-6947-10-14).

[26] Höhle M, Mazick M. Aberration detection in R illustrated by Danish mortality monitoring. In: Kass-Hout, T. and Zhang, X. (editors). *Biosurveillance: Methods and Case Studies, CRC Press* 2010: 215–238.

[27] Robertson C, Nelson TA. Review of software for space-time disease surveillance. *International Journal of Health Geographics* 2010; **9**:16, (online;doi: 10.1186/1476-072X-9-16).

[28] Hulth A, Andrews N, Ethelberg S, Dreesman J, Faensen D, van Pelt W, Schnitzler J. Practical usage of computer-supported outbreak detection in five European countries. *Eurosurveillance* 2010; **15**, (online; doi:pii: 19851).