Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713428038

Multivariate outbreak detection

Linus Schiöler[a]; Marianne Frisén[a]
[a] Statistical Research Unit, Department of Economics, University of Gothenburg, Gothenburg, SE, Sweden

First published on: 25 May 2011

To cite this Article Schiöler, Linus and Frisén, Marianne(2011) 'Multivariate outbreak detection', Journal of Applied Statistics,, First published on: 25 May 2011 (iFirst)
To link to this Article: DOI: 10.1080/02664763.2011.584522
URL: http://dx.doi.org/10.1080/02664763.2011.584522

Journal of Applied Statistics
2011, iFirst article

Taylor & Francis
Taylor & Francis Group

Multivariate outbreak detection

Linus Schiöler and Marianne Frisén*

Statistical Research Unit, Department of Economics, University of Gothenburg, Gothenburg, SE 40530, Sweden

(*Received 10 October 2010; final version received 18 April 2011*)

Online monitoring is needed to detect outbreaks of diseases such as influenza. Surveillance is also needed for other kinds of outbreaks, in the sense of an increasing expected value after a constant period. Information on spatial location or other variables might be available and may be utilized. We adapted a robust method for outbreak detection to a multivariate case. The relation between the times of the onsets of the outbreaks at different locations (or some other variable) was used to determine the sufficient statistic for surveillance. The derived maximum-likelihood estimator of the outbreak regression was semi-parametric in the sense that the baseline and the slope were non-parametric while the distribution belonged to the one-parameter exponential family. The estimator was used in a generalized-likelihood ratio surveillance method. The method was evaluated with respect to robustness and efficiency in a simulation study and applied to spatial data for detection of influenza outbreaks in Sweden.

Keywords: exponential family; generalized likelihood; ordered regression; spatial data; surveillance

1. Introduction

Online surveillance is used to give an alert signal as soon as possible after an important change has occurred. Overviews of the inferential issues in surveillance are given in [8,9,25,34,43] and others.

Here, we will consider the detection of an outbreak, defined as a change from a (possibly unknown) baseline to a monotonically increasing (or decreasing) regression. Other definitions of outbreaks are discussed in Section 7.

The motive for this study was the spatial surveillance of influenza outbreaks. The detection of outbreaks of epidemiological diseases is an important area of online surveillance. Surveillance in public health is reviewed in, e.g. [23,26,37,39,42]. By monitoring incidences, outbreaks of reoccurring diseases may be detected, for example, the yearly influenza epidemic. Such monitoring is also useful to detect new diseases, such as SARS, avian flu and swine influenza, as well as effects of bioterrorism. Early detection of the onset of an outbreak is useful in order for health authorities to act timely and also for the planning of health care. Epidemics, such as influenza, are for several

*Corresponding author. Email: marianne.frisen@statistics.gu.se

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664763.2011.584522
http://www.informaworld.com

reasons very costly to society and it is therefore of great value to monitor the epidemic period in order to properly allocate medical resources [2]. A semi-parametric method for detecting the onset of a monotonic increase was suggested for univariate surveillance by Frisén and Andersson [10]. It was successfully applied to the incidence of influenza in Sweden as a whole in [12].

As information on the incidence in different regions of the country is available, we will here generalize the univariate method to utilize this information. Spatial surveillance is a special case of multivariate surveillance, as pointed out, for example, in [22,40]. The relation between different variables (here locations) is important in the monitoring of the onset of the outbreak. We will use information from a study in [35] on the spread of influenza in Sweden. The spreading pattern is described in Section 6.1. We will investigate how information on time lags in the onset at different locations should be used in an outbreak surveillance system. Another case where a time lag might be relevant is when you have an early but rough indicator which might be combined with a later and more accurate one. In [16,20], it was shown that data of search patterns on the Internet could be used as a proxy for influenza incidence. In [16], it was found that the lag in reporting was about 1 day compared with between 1 and 2 weeks for traditional data from the U.S. Centers for Disease Control and Prevention. The method suggested in this paper may possibly be useful also for situations like that one, where the lag is in the reporting rather than in the onset of the outbreak at various locations.

In Section 2, we will specify univariate and multivariate models for outbreaks in order to clarify which changes we aim to detect. In Section 3, we will derive a sufficient reduction in the data for multivariate outbreak situations. This reduces the complexity without loss of information. Sufficient reduction for detection of step changes was earlier derived in [14] but here it is derived for the detection of gradual outbreaks. In Section 4, we will discuss general approaches of how multivariate surveillance can be constructed from univariate surveillance and construct a simple multivariate outbreak detection method, based on the univariate method in [10]. In this section, we will also derive the recommended method. This is done by deriving the maximum-likelihood estimators based on the multivariate monotonicity restrictions and using these in a generalized likelihood ratio (GLR) method. In Section 5, we evaluate the suggested method by a simulation study, where properties such as predictive value and robustness are examined. The robustness is important since you never can expect assumptions to be exactly fulfilled. In the comparison with other methods, we will use the evaluation metrics suggested by Frisén *et al*. [13] for multivariate surveillance. In Section 6, the method is applied to data for several influenza seasons in Sweden, and the efficiency of the suggested multivariate outbreak detection method is demonstrated. Concluding remarks are given in Section 7.

2. Specification of the outbreak model

Since the method for outbreak detection depends on the specification of the model, we start with this specification. At each time point, t, a new observation is made on a process $\mathbf{Y}$. We state the model for discrete time. Weekly data are available for influenza in Sweden. We want to detect the change from one state to another as soon as possible after it has occurred, in order to give warnings and to take corrective actions.

2.1 *Univariate outbreak*

In [1], Swedish influenza data from six seasons (2001–2007) were analyzed, and it was suggested that a non-parametric approach based on monotonicity restrictions (the outbreak regression) should be used. It was also suggested that the outbreak could be modeled using a Poisson distribution for the incidence. The localization parameter $\lambda(t)$ of the distribution at time t has a constant value λ_0 before the outbreak but depends on time after the onset of the outbreak. We will use τ to

denote the unknown time of the onset. Thus,

$$\lambda(t) = \begin{cases} \lambda_0, & t < \tau \\ \lambda_{t-\tau+1}, & t \geq \tau, \end{cases}$$

with $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_s$. This situation where the regression is constant at first and then monotonically increasing will be named "outbreak regression". The aim at decision time s is to determine whether or not the outbreak has started yet, thus if $\tau \leq s$ or $\tau > s$. The state at the outbreak is characterized by a monotonically increasing expected incidence.

2.2 *Multivariate outbreak*

In multivariate surveillance, the process under surveillance is a p-variate vector, denoted by $\mathbf{Y} = \{\mathbf{Y}(t), \quad t = 1, 2, \ldots\}$, where $\mathbf{Y}(t) = \{Y_1(t), Y_2(t), \ldots, Y_p(t)\}$. The components of the vector represent, for example, the incidence of a disease at different locations. For the influenza in Sweden, described in Section 6, we have $p = 2$. Each component $Y_i(t)$ is distributed with a location parameter $\lambda(t)$ with the same properties as described in Section 2.1. The time of the onset may differ for the components and will be denoted τ_i for component i. At decision time s, we base the decision whether an outbreak has occurred or not on the available information, $\mathbf{Y^s} = \{\mathbf{Y}(1), \mathbf{Y}(2) \cdots \mathbf{Y}(s)\}$.

The time τ_i of the onset of the outbreak of process Y_i may not be the same for all $i = 1, \ldots, p$. The relation between the times is important since totally different methods will be optimal for different relations. The aim here is to detect an outbreak in any of the processes, which means that we aim at detecting the first one. We will concentrate on the case of a known time lag. This can be the case for spatial data and data from several sources (possibly including proxy data). The case where the lag is mis-specified is examined in Section 5.5. Since the time lags are assumed to be known, for notational convenience, we will order the processes according to which changes first. Hence, $\tau_1 \leq \cdots \leq \tau_p$. The time lag for process Y_i will be denoted by q_i, where $q_i = \tau_i - \tau_1$ for $i = 1, \ldots, p$. The case where the onsets are simultaneous, that is, $\tau_i = \tau$ for $i = 1, \ldots, p$, is of special interest. In this case, $q_i = 0$, $i = 1, \ldots, p$. We denote this by lag $= 0$. In numerical examples and applications, we will also use the special cases of only two processes with $q_2 = 1$ or 2. We denote this by lag $= 1$ and lag $= 2$, respectively. For the influenza in Sweden, lag $= 0$, which is a total over the country, has previously been used. In Section 6, we examine possible benefits of combining different parts of the country with lag $= 1$.

We assume that the distributions of the processes all belong to the one-parameter exponential family. In the application to influenza data in Section 6, the Poisson distribution is relevant. We do not assume a parametric outbreak pattern here. Instead, we assume that the different processes are identically distributed, except for the time of the onset.

In order to derive a good method for outbreak detection for the model described in this section, we will first examine possibilities to reduce the data without loss of information. In Section 3, we will explore the possibility of such a sufficient reduction.

3. Sufficient reduction at multivariate outbreaks

Due to the complexity of multivariate problems, we will now examine the possibilities, for outbreak detection, to minimize the complexity without loss of information. A sufficient reduction will not reduce the information and still allows a joint solution to the full surveillance problem. In [13], it was demonstrated that the relation between the change points affects the properties of different surveillance methods in different ways. It is thus necessary to consider any knowledge on the relation between changes. Without any information about this relation, it is not possible to

derive a good method. In most papers on multivariate surveillance, it is implicitly assumed that the changes occur simultaneously. It is thus of special interest to study a simultaneous outbreak at all locations and its implications. We will also study a time lag in the onset of the outbreaks. This situation is of interest for influenza in Sweden as described in Section 6. Robustness when the time lag is only approximately known is studied in Section 5.5.

3.1 *Simultaneous change at all locations*

Many evaluations of multivariate surveillance methods are made by the zero-state average run length (ARL) (see Section 5.3), where the change occurs at the start. When all processes change at the start, it follows that they change simultaneously.

In [14,41], it was demonstrated that if all processes have the same change points, i.e. $\tau_1 = \tau_2 = \cdots = \tau_p = \tau$, then the univariate vector of partial-likelihood ratios, $\{L(s,t), t = 1, \ldots, s\}$ where $L(s,t) = f(Y; \tau = t \leq s)/f(Y; \tau > s)$, is sufficient for the sequence of distributional families. Thus, in order to monitor a simultaneous fully specified change in distribution, it is possible to construct a univariate surveillance procedure based on the sufficient sequence of likelihood ratios. This result was used in [45] for the simultaneous shifts of mean and variance in a normal distribution. For the case with no lag between the change points of two processes (lag $= 0$), the sufficient statistic is denoted by SuffR0. We will use this notation in the application of spatial surveillance of Swedish influenza outbreaks. In this case, SuffR0 corresponds to the total incidence in the country as a whole. The statistic OutbreakPSuffR0 of the method in the application is hence equivalent to the statistic of the univariate surveillance of influenza in Sweden reported in [10,12].

3.2 *Changes with a time lag between locations*

The case of a known time lag for independent normal distributions with equally sized shifts in the expected value at the change points was studied in [21], where it was demonstrated that a sufficient reduction to univariate surveillance exists. In [14], the case of changes in the general one-parameter exponential family was studied, but also only for step changes. Different levels of the parameter before the change as well as differences in shift size were considered. These earlier results on sufficiency are only valid for step changes, and hence inadequate for our aim to detect changes from a constant level to a monotonically increasing one. Theorem 1 shows that a sufficient reduction to a univariate statistic exists for the situation where each process Y_i increases monotonically from the onset of the outbreak τ_i and onwards, and there is a known time lag between the onsets of each process. The indices of the observation vectors $\{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_p\}$ are ordered according to ascending time lag, i.e. the change occurs first in $\mathbf{Y}_1$. The theorem is illustrated for a simple case in Example 1 (after Theorem 1). A numerical illustration is given in Example 2 in Section 4.6.

THEOREM 1 *For p processes $Y_1, Y_2, \ldots, Y_p$ which all belong to the one-parameter* (*for localization*) *exponential family and which are independent and identically distributed*, *conditional on the change points and time lags* (*independent over time as well as across processes*), *there exists a sufficient reduction of the set of observation vectors to a univariate statistic for the detection of outbreaks with equal* (*but possibly unknown*) *parameter values from the onset of the outbreak when the changes occur with known time lags* $(q_1 = 0, q_2, q_3, \ldots, q_p)$ *where* $q_i = \tau_i - \tau_1$. *A sufficient statistic for inference on the first onset τ_1 is the sequence*

$$\sum_{i \in I_t} Y_i(t + q_i), t = 1, \ldots, s, \quad \text{where } I_t = \{i : q_i \leq s - t, 1 \leq i \leq p\}.$$

This is true both for the situation when the time of change is fixed but unknown and for a stochastic time of change.

The proof is given in Appendix 1.

Since any one-to-one function of a sufficient statistic also is sufficient, the sequence

$$\sum_{i\in I_t}\frac{Y_i(t+q_i)}{|I_t|},\quad t=1,\ldots,s,$$

where $|I_t|$ denotes the cardinality of I_t, is sufficient. This transformed statistic is useful when dealing with the monotonicity restrictions of the outbreak regression, since this statistic preserves the monotonicity properties.

When we have two processes, we will use a simpler notation, $\text{SuffR}q(s,t)=\sum_{i\in I_t}Y_i(t+q_i)/|I_t| : t=1,\ldots,s$, where $q=q_2$ is the lag between the two processes.

Example 1 For two processes Y_1 and Y_2 with time lag $q=1$, the index set is $I_t=\{i : q_i\le s-t, 1\le i\le p\}$. For $s=1$, we have $I_1=\{i : q_i\le 0, 1\le i\le 2\}=\{1\}$. For $s=2$, we have $I_1=\{i : q_i\le 1, 1\le i\le 2\}=\{1,2\}$ and $I_2=\{i : q_i\le 0, 1\le i\le 2\}=\{1\}$. For $s=3$, we have $I_1=\{i : q_i\le 2, 1\le i\le 2\}=\{1,2\}$, $I_2=\{i : q_i\le 1, 1\le i\le 2\}=\{1,2\}$ and $I_3=\{i : q_i\le 0, 1\le i\le 2\}=\{1\}$. Hence, the sufficient reduction is $\left\{\sum_{i=1}Y_i(t) : t=1\right\}=\{Y_1(1)$ at $s=1$, $\left\{\sum_{i\in I_t}Y_i(t+q_i) : t=1,2\right\}=\left\{\sum_{i\in\{1,2\}}Y_i(1+q_i),\sum_{i\in\{1\}}Y_i(1+q_i)\right\}=\{Y_1(1)+Y_2(2),Y_2(2)\}$ at $s=2$, $\{Y_1(1)+Y_2(2),\ Y_1(2)+Y_2(3),\ Y_1(3)\}$ at $s=3$ or more generally $\{Y_1(1)+Y_2(2),Y_1(2)+Y_2(3),\ldots,Y_1(s-1)+Y_2(s),Y_1(s)\}$ at s. A numerical example is given in Section 4.6.

The sufficient statistic at decision time s is $\text{SuffR}q(s,t), t=1,\ldots,s$, where $\text{SuffR}q(s,t)=(Y_1(t)+Y_2(t+q))/2$ for $t\le s-q$ and $\text{SuffR}q(s,t)=Y_1(t)$ for $t>s-q$. In Example 1, we have $\{\text{SuffR1}(1,t)\}=\{\{Y_1(1)\}$ at $s=1$. At $s=2$, we have $\{\text{SuffR1}(2,t)\}=\{[Y_1(1)+Y_2(2)]/2,Y_2(2)\}$. At $s=3$, we have $\{\text{SuffR1}(3,t)\}=\{[Y_1(1)+Y_2(2)]/2,[Y_1(2)+Y_2(3)]/2,Y_1(3)\}$. More generally, we have $\{\text{SuffR}p(p,t)\}=\{[Y_1(1)+Y_2(2)]/2,\ldots,[Y_1(2)+Y_2(3)]/2,\ldots,[Y_1(s-1)+Y_2(s)]/2,Y_1(s)\}$.

4. Surveillance methods for multivariate outbreak detection

In this section, we will first describe the univariate outbreak detection method, OutbreakP, suggested by Frisén and Andersson [10]. Then, we will review common approaches to adapting univariate surveillance to multivariate surveillance and show how OutbreakP can be adapted by these approaches. After that, we will derive a joint multivariate method based on the sufficiency principle. Finally, we will give the maximum-likelihood estimator of the parameters and a GLR method for outbreak detection.

4.1 *Univariate outbreak detection*

For the outbreak detection situation, one way to specify the in-control state versus the outbreak is to use a parametric model of the outbreak curve. This requires extensive modeling as in, e.g. [17]. Here, we will use a non-parametric univariate method as a base for the suggested adaption to a multivariate situation. When seasonal or other components are important, it might be useful to apply the non-parametric method to the residuals of a more complex model.

As we have unknown parameters, GLR where parameters are substituted with the maximum-likelihood estimates is used. This general approach was introduced to surveillance in [25], where

it was suggested that in the CUSUM method, GLR should be used to handle unknown parameters after the change. This approach was also used by Höhle and Paul [19] for Poisson and negative binomial distribution at surveillance of infectious diseases. In [10], our method for outbreak detection was suggested. The method utilized the GLR approach by using the maximum-likelihood estimators under the monotonicity restrictions in Section 2.1, as derived in [11] for the one-parameter exponential family. The method was derived for the normal and Poisson distributions and was named the OutbreakP method for the Poisson distribution. Here, we will only consider the Poisson distribution, which is suitable for the application in Section 6. However, the same technique was used for a normal distribution with known variance in [11] for the univariate case. The method is semi-parametric since the distribution is parametric, but the regression is non-parametric since the only restriction on the regression on the parameter of the one-parameter exponential family is by monotonicity. A user-friendly computer program can be downloaded at www.statistics.gu.se/surveillance. The method is also available in the R package Surveillance, described in [18] and available on CRAN, and the open JAVA package CASE described in [4].

For the univariate surveillance of the influenza incidence in Sweden as a whole, the OutbreakP method was evaluated in [10,12]. We will now adapt this method for a multivariate situation.

4.2 *General approaches to adapting univariate surveillance to multivariate surveillance*

There are several approaches to multivariate surveillance. The most commonly used approach is the reduction to one scalar statistic, such as the sum for each time. This will be described in Section 4.3. Another approach is to use several univariate systems in parallel, one for each process. An intermediate approach is vector accumulation, for example, MEWMA suggested by Lowry *et al*. [28]. When the multivariate distribution is available, as in, e.g. [30], this might be used as a base for a surveillance method [3,40].

4.3 *Reduction to one scalar statistic for each time*

Dimension reduction is always a reasonable choice in multivariate problems provided that it does not reduce important information. The most far-going reduction is the reduction to a scalar for each time. This is the most common way to handle multivariate surveillance. The observations at each time point consist of a vector, and we can first transform the vector from the current time point into a scalar statistic, which we then accumulate over time.

One example of scalar accumulation is when, for each time point, a statistic representing the spatial pattern is constructed. This statistic is then used in a surveillance method. The reduction to a univariate variable can be followed by univariate monitoring of any kind. In [32,33], statistics measuring clustering were used for each time, and the information was accumulated by the univariate CUSUM method.

In [44], the spatial pattern was characterized by a Bayesian model for each time, and the statistic was then monitored by the EWMA method.

For the influenza incidence, a commonly used reduction is the sum, even though information on different parts of the country is available. Using the sum means that no regional information is used. Instead, the surveillance is based on total data for the country as a whole, as in [10]. However, other reductions may be more efficient, as is seen in Section 3. In our evaluations in Section 5, the reduction to a scalar is included.

4.4 *Parallel outbreak detection*

To illustrate a frequently used approach to multivariate surveillance, we will include a parallel system in our evaluations. Here, each process is monitored separately and an overall alarm is

called if some condition is fulfilled. The most common condition is that one of the systems calls an alarm. We will use this condition when the univariate OutbreakP method is applied to each process. The method is called OutbreakPParallel. Results for this method, when compared with others, are given in Section 5.3.

4.5 *Outbreak surveillance based on sufficient reduction and known parameters*

The partial-likelihood ratio of an outbreak versus no outbreak with onsets of the outbreaks at τ_1, $\tau_2, \ldots, \tau_p$ is

$$L(s, t_1, \ldots, t_p) = \frac{f(\mathbf{Y}^s | \tau_1 = t_1, \ldots., \tau_p = t_p)}{f(\mathbf{Y}^s | \tau_1 > s, \ldots, \tau_p > s)}.$$

For known time lags $(q_1 = 0, q_2, q_3, \ldots, q_p)$, this can be written as

$$L(s, t_1) = \frac{f(\mathbf{Y}^s | \tau_1 = t_1)}{f(\mathbf{Y}^s | \tau_1 > s)}.$$

For the detection of an outbreak as defined in Section 2, $L(s, 1)$ is the relevant statistic [10]. For the Poisson distribution and known values of the parameters of the regressions, we have that

$$L(s,1) = \prod_{i=1}^{p} \prod_{t=1+q_i}^{s} \exp(\lambda_0 - \lambda_{t-q_i}) \left(\frac{\lambda_{t-q_i}}{\lambda_0}\right)^{Y_i(t)} = \prod_{t=1}^{s} e^{|I_t|(\lambda_0 - \lambda_t)} \left(\frac{\lambda_t}{\lambda_0}\right)^{\sum_{i \in I_t} Y_i(t+q_i)},$$

where $I_t = \{i : q_i \leq s - t, 1 \leq i \leq p\}$.

For two processes, we have

$$L(s,1) = \prod_{t=1}^{s-q} \mathrm{e}^{2(\lambda_0 - \lambda_t)} \left(\frac{\lambda_t}{\lambda_0}\right)^{Y_1(t) + Y_2(t+q)} \prod_{t=s-q+1}^{s} \mathrm{e}^{\lambda_0 - \lambda_t} \left(\frac{\lambda_t}{\lambda_0}\right)^{Y_1(t)}.$$

In Section 4.7, we will use the generalized maximum likelihood and substitute the unknown parameters with their maximum-likelihood estimators derived in Section 4.6.

4.6 *Maximum-likelihood estimation of the multivariate outbreak regression*

If the distribution of the processes is not fully specified, the approach of the GLR can be used. Hence, we need estimates for the likelihood ratio in Section 4.5, both for the situation with an outbreak and for the situation with no outbreak. When we have no outbreak, and thus all observations are independent and identically distributed, the maximum-likelihood estimator of λ_0 is the average of all observations. We have

$$\hat{\lambda}_0 = \sum_{t=1}^{s} \sum_{i=1}^{p} \frac{y_i(t)}{sp}.$$

In the outbreak situation, we have the monotonicity restriction described in Section 2. A useful technique to find least-squares estimates, which here are maximum-likelihood estimates, is the

pool adjacent violator algorithm (PAVA), described, for example, by Robertson *et al*. [31]. This algorithm was introduced to surveillance in [5]. However, both their models and aim of the surveillance differ from ours.

THEOREM 2 *For the multivariate outbreak regression in Section* 2.2 *with processes which all belong to the regular one-parameter* (*for localization*) *exponential family and which are independent and identically distributed*, *conditional on the change points and known time lags* (*independent over time as well as across processes*), *the maximum-likelihood estimators of* λ_t, *for the increasing phase are obtained by the PAVA algorithm with weights proportional to the number,* $|I_t|$, *of processes used for the specific component of the sufficient statistic*.

The proof is given in Appendix 2.

Example 2 To illustrate how the sufficient reduction and the PAVA algorithm are used, we give a simple example for two processes with lag $q = 1$. SuffR$q(s, t)$ is the sufficient reduction described in Section 3.2, where q indicates the lag between the two processes and s is the decision time. In Table 1, we illustrate how the sufficient statistic and the maximum-likelihood estimators are calculated for a numerical example.

The estimate of $\hat{\lambda}_0$ is the average of all observations. At $s = 5$, we have. To estimate $\hat{\lambda}_t$ at time $s = 5$, we apply the PAVA to the sequence SuffR1(5,t), $t = 1, \ldots, 5$. We see that the first violation of the order restriction occurs at $t = 2$, and hence we replace the observations by the weighted average, $(2.5 \cdot 2 + 2 \cdot 2)/4 = 2.25$. This does not violate the first observation, $Y_2(1)$, since $2 \leq 2.25$. The observation at $t = 4$ constitutes a violation, and hence we use $(3 \cdot 2 + 1.5 \cdot 2)/4 = 2.25$, which does not violate the order restriction of the previous observations.

4.7 *GLR surveillance of multivariate outbreaks*

We will use the GLR, i.e. substitute parameter values by their maximum-likelihood estimators, in our semi-parametric multivariate method.

By substituting the parameters of the outbreak regression in $L(s, 1)$ in Section 4.5 with the maximum-likelihood estimators in Section 4.6, we get the alarm statistic of the multivariate OutbreakPSuffR method. Here P stands for the Poisson distribution while SuffR stands for the sufficient reduction in the multivariate case. The general method depends on the set of lags $(q_1 = 0, q_2, q_3, \ldots, q_p)$ and has the alarm statistic

$$\prod_{i=1}^{p} \prod_{t=1+q_i}^{s} \exp(\hat{\lambda}_0 - \hat{\lambda}_{t-q_i}) \left(\frac{\hat{\lambda}_{t-q_i}}{\hat{\lambda}_0} \right)^{Y_i(t)} = \prod_{t=1}^{s} e^{|I_t|(\hat{\lambda}_0 - \hat{\lambda}_t)} \left(\frac{\hat{\lambda}_t}{\hat{\lambda}_0} \right)^{\sum_{i \in I_t} Y_i(t+q_i)},$$

Table 1. For an example of observations on two processes, we give the sufficient statistic SuffR1 for $s = 1$, 2, 3, 4, 5 and the maximum-likelihood estimate $\hat{\lambda}_t$ at $s = 5$.

t	y_1	Y_2	SuffR1(1,t)	SuffR1(2,t)	SuffR1(3,t)	SuffR1(4,t)	SuffR1(5,t)	$\hat{\lambda}_t$
1	4	2	4	2.5	2.5	2.5	2.5	2.25
2	3	1		3	2	2	2	2.25
3	3	1			3	3	3	2.25
4	1	3				1	1.5	2.25
5	6	2					6	6

where $I_t = \{i : q_i \leq s - t, 1 \leq i \leq p\}$. For two processes with time lag q, we use the notation OutbreakPSuffRq for the method and OutbreakPSuffR$q(s)$ for the alarm statistic. For this case, we have

$$\prod_{t=1}^{s-q} \mathrm{e}^{2(\hat{\lambda}_0 - \hat{\lambda}_t)} \left(\frac{\hat{\lambda}_t}{\hat{\lambda}_0}\right)^{Y_1(t)+Y_2(t+q)} \prod_{t=s-q+1}^{s} \mathrm{e}^{\hat{\lambda}_0 - \hat{\lambda}_t} \left(\frac{\hat{\lambda}_t}{\hat{\lambda}_0}\right)^{Y_1(t)}.$$

In the case $q = 0$, this simplifies to the univariate OutbreakP statistic described in [10,12].

Example 3 For the situation of Examples 1 and 2, we have for $s = 5$ the alarm statistic

$$\text{OutbreakPSuffR1(5)} = \prod_{t=1}^{4} \mathrm{e}^{2(\hat{\lambda}_0 - \hat{\lambda}_t)} \left(\frac{\hat{\lambda}_t}{\hat{\lambda}_0}\right)^{Y_1(t)+Y_2(t+1)} \prod_{t=5}^{5} \mathrm{e}^{\hat{\lambda}_0 - \hat{\lambda}_t} \left(\frac{\hat{\lambda}_t}{\hat{\lambda}_0}\right)^{Y_1(t)} = 6.14.$$

5. Simulation study to determine the properties of the multivariate OutbreakP method

In a multivariate situation, some reduction in the dimensionality of data is often useful, but it is important that no information is lost. This could be achieved by the use of a sufficient statistic. If the outbreaks appear simultaneously for the different processes, then we have a univariate sufficient statistic with one change point. However, when the outbreaks appear at different times, the sufficient statistic has more than one change point in the distribution. Even though each component has one change point, the distribution of the sufficient statistic is not constant either for $t < \tau_i$ or for $t \geq \tau_i$. The proofs commonly used for minimax or expected delay optimality require that there is only one change between two distributions.

Since exact optimality cannot be expected, the properties of the OutbreakP method are presented by the results from a simulation study. In Section 6, the method will be evaluated by the application of the method to observed Swedish influenza data.

5.1 *Model for simulations*

The suggested method is non-parametric with respect to the shape. However, to examine the properties of the method by a simulation study, a parametric model was used to generate data. We used a model that is relevant for the application to the influenza data described in Section 6. In the model, we have two independent processes, Y_1 and Y_2, generated as

$$Y_i(t) \sim \text{Poisson}(\lambda_i(t))$$

$$\lambda_i(t) = \begin{cases} 0.5 & t < \tau_i \\ \exp(\beta_0 + \beta_1(t - \tau_i + 1) & t \geq \tau_i, \end{cases}$$

where $\beta_0 = -0.622$ and $\beta_1 = 0.826$. In accordance with Section 2.2, we have $\tau_{\min} = \tau_1$ and $\tau_2 = \tau_1 + q$, where q denotes the time lag. For each value of and t, we generated 10^6 replicates.

The model is based on the studies in [1,35] on the seasonal influenza in Sweden. The parameters were estimated from Swedish influenza data from the season 2003–2004, which was not extreme in any sense but "typical". This model also fits rather well for many seasons; see [35] for further details.

5.2 *False alarms*

The most commonly used measure for false alarms is the in-control, ARL^0, $E[t_\mathrm{A}|\tau = \infty]$. This can be used also in a multivariate situation. A similar measure is the median run length, MRL^0. Because of the skewness of the run-length distribution, the median might be easier to interpret. In addition, considerably less computer time is necessary for the same accuracy, since good estimates of the low frequency of long runs are needed for ARL^0 but not for MRL^0. We determined the alarm limits so that we have the same MRL^0 (780) in all comparisons in this paper. It was used also for the univariate OutbreakP method in [10]. The technique chosen by Frisén and Sonesson [15] was used to ensure that the alarm limit was determined with enough accuracy to make the error in the curves of delay less than the line width.

5.3 *Delay*

One measure of the detection ability is the average run length, given that the change occurs immediately ($\tau = 1$). This is widely used in univariate surveillance and often named zero-state ARL or ARL^1. Zero-state ARL is the most commonly used evaluation measure also in the multivariate case. However, it is seldom explicitly defined. The definition implicit in most publications is $E[t_\mathrm{A}|\tau_1 = \tau_2 = \cdots \tau_p = 1]$, where t_A is the time of the alarm. Here, it is assumed that all processes change at the same time. As seen in Section 3.1, a sufficient reduction to a univariate problem exists when all processes change at the same time. Zero-state ARL is thus questionable as a formal measure for comparing methods for genuinely multivariate problems. Instead, we will here use a measure which allows different change points.

The conditional expected delay $\mathrm{CED}(\tau) = E[t_\mathrm{A} - \tau|\tau \leq t_\mathrm{A}]$ can be generalized for multivariate surveillance to $\mathrm{CED}(\tau_1, \tau_2, \ldots, \tau_p) = E[t_\mathrm{A} - \tau_{\min}|\tau_{\min} \leq t_\mathrm{A}]$ [13]. For a given lag, this depends on only one of the change points. Thus, we can write $\mathrm{CED}(\tau_{\min}) = E[t_\mathrm{A} - \tau_{\min}|\tau_{\min} \leq t_\mathrm{A}]$. When we have lag $= 0$, i.e. simultaneous outbreaks, this reduces to the univariate CED. In Figure 1, we can see that the OutbreakPParallel method has a worse delay than the OutbreakPSuffR0 method for simultaneous outbreaks. OutbreakPSuffR0 is based on SuffR0, which corresponds to the total incidence. In Figure 2, with Monte Carlo estimates, we can see that the

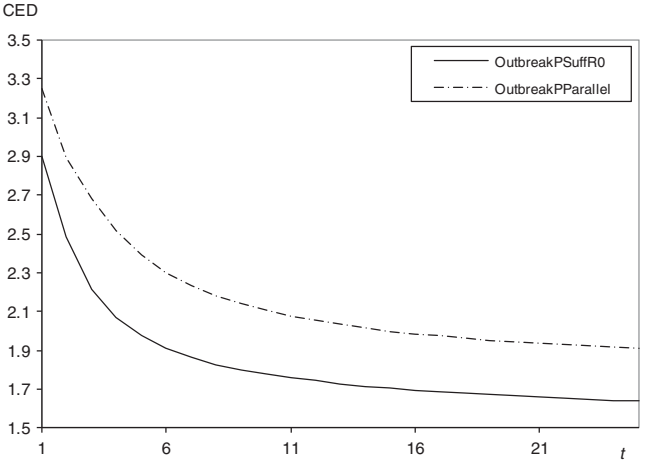

Figure 1. The CED for the OutbreakPParallel and OutbreakPSuffR0 methods for two processes with simultaneous onset of the outbreak (lag $= 0$) as a function of $\tau_{\min} = t$.

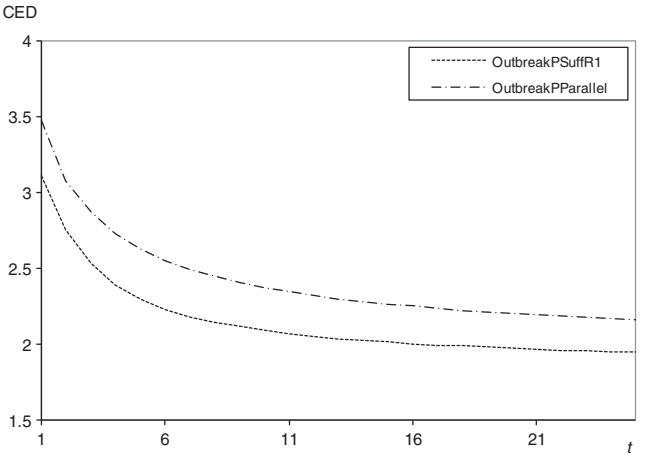

Figure 2. The delay in the detection of the outbreak for the OutbreakPParallel and OutbreakPSuffR1 methods for two processes with lag $= 1$ as a function of $\tau_{\min} = t$.

delay for the parallel method is worse than that for the OutbreakPSuffR1 method based on SuffR1 when lag $= 1$.

5.4 *Predictive value*

If a method calls an alarm, it is important to know whether this alarm is a strong or weak indication of a change. The predictive value is a well-established measure in epidemiology. In surveillance, however, we need a variant that also incorporates time. The difference in surveillance, when compared with situations involving only one decision, is that we can get an alarm at any time point, and therefore we need a measure of the predictive value at each of them. In order to judge to what degree an alarm at time t_A can be trusted, it is necessary to consider the balance between the risk of false alarms, the detection ability and the probability of a change. This can be done by calculating the probability of an outbreak, at an alarm:

$$\begin{aligned} \mathrm{PV}(t) &= P(\tau_{\min} \le t | t_A = t) \\ &= \frac{\sum_{i=1}^{t} (P(t_A = t | \tau_{\min} = i) P(\tau_{\min} = i))}{\sum_{i=1}^{t} (P(t_A = t | \tau_{\min} = i) P(\tau_{\min} = i)) + P(t_A = t | \tau_{\min} > t) P(\tau_{\min} > t)} \end{aligned}$$

as suggested in [7,13]. The components in the formula depend in general on the relation between the change points. Here, the known lags have been used to simplify the calculations. The predictive value depends also on whether outbreaks appear frequently or rarely. Knowledge of the exact distribution of $\tau_{\min}$ is seldom available, but we will nevertheless try to give a rough indicator. In the simulation study, $\tau_{\min}$ was assumed to be geometrically distributed, i.e. $P(\tau_{\min} = i) = (1 - \nu)^{i-1}\nu$. This may not give the closest fit of the onset times in Sweden, but in order to detect outbreaks which occur at unexpected times we did not want to include information on which week is the most common one for the onset. The level of intensity was roughly estimated from all available historical data on seasonal influenza to be $\nu = 0.1$. This value is thus consistent with the application in Section 6. With this intensity, the PV is above 0.99, and for a lower intensity, $\nu = 0.01$, which weakens the PV, it is above 0.95. The method and alarm limit used in the simulation study were considered potentially useful for practical application since the predictive value was high.

5.5 *Robustness to mis-specified time lag*

Some models and assumptions are needed in order to efficiently make inferences from data. Hence, it is important to make assumptions which are suitable for the application. Here, we will concentrate on robustness related to a possible time lag. First, we will describe the effect of using the method but with a wrong lag, then we will describe the consequences of different population sizes of different regions.

The lag between the outbreaks is seldom exactly known. For situations of interest for the application in Section 6, we examine the effect of using the sufficient statistic for lag $= 1$ when in fact lag $= 2$, and vice versa and we also examined the case where the true lag is between 1 and 2. For comparison, we also included the use of a statistic based on a lag further away from the true one. In Figure 3, we have simulated influenza outbreaks where the true lag is 1. We can see that when we used the method OutbreakPSuffR1, which is based on the true lag, we got the best results. When we used the method for lag $= 2$ or lag $= 0$, the results were slightly worse while it is much worse when we used the method for lag $= 4$. In Figure 4, we report simulated outbreaks where the true lag is 1.5. Here, we can see that the methods based on lag $= 1$ and lag $= 2$ both works fine while those for lag $= 0$ and lag $= 4$ are clearly worse. In Figure 5, we have simulated outbreaks with the true lag 2. When we used the outbreak detection method based on the true lag, we got the best results, except for a very minor advantage for SuffR1 at $\tau = 1$ and 2. In this complex situation, the method based on the sufficient statistic is not always exactly optimal, but it usually works very well. When we used the statistic for lag $= 1$, the results were similar to those for the true lag. However, when the lag was two steps away from the true one and we used the sufficient statistic for lag $= 0$ or lag $= 4$, while the true lag was 2, we got clearly worse results. The conclusion is that an approximate lag may work well, for the case of the application of Section 6, provided that it is not too far away from the true one. Improvements from the present use of lag $= 0$ seem possible.

In the simulation model used above, we assumed equal distributions given the possibly different times of onset. In practice, however, the two processes may be based on different population sizes or otherwise have different parameters. If the difference is large, this should be handled by the adjustment of the weights and the alarm limit. The ratio in size between the two areas analyzed in Section 6 is approximately 1.17, and a suitable simulation model for this case was derived in [35]. We examined what would happen if no adjustments were made and the same weights and

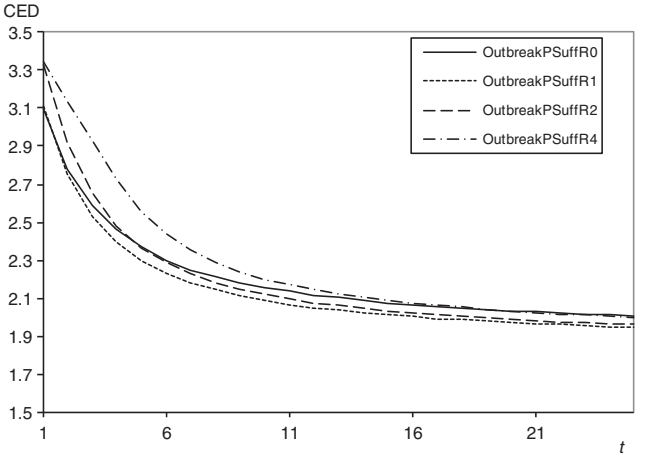

Figure 3. The delay, as a function of $\tau_{\min} = t$, for outbreak detection by OutbreakPSuffR0, OutbreakPSuffR1, OutbreakPSuffR2 and OutbreakPSuffR4 when the true lag is 1.

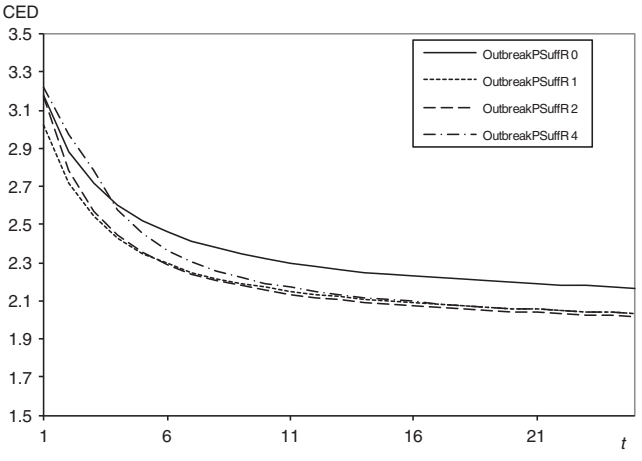

Figure 4. The delay, as a function of $\tau_{\min} = t$, for outbreak detection by OutbreakPSuffR0, OutbreakPSuffR1, OutbreakPSuffR2 and OutbreakPSuffR4 when the true lag is 1.5.

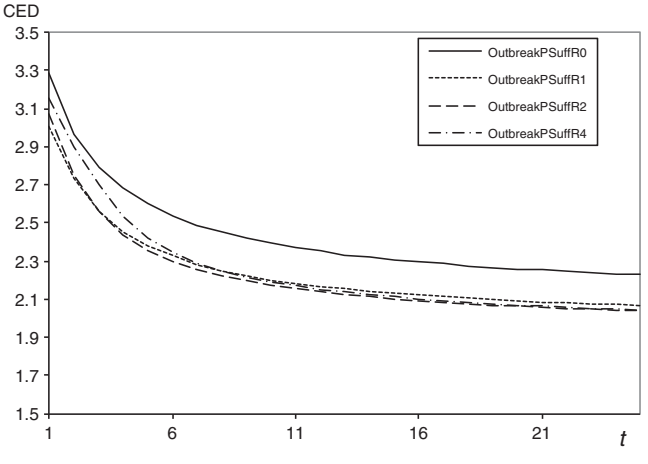

Figure 5. The delay, as a function of $\tau_{\min} = t$, for outbreak detection by OutbreakPSuffR0, OutbreakPSuffR1, OutbreakPSuffR2 and OutbreakPSuffR4 when the true lag is 2.

alarm limit were used, as if the population sizes were the same. The OutbreakPSuffR methods performed slightly worse if different population sizes were used. However, the predictive value of an alarm was still greater than 0.99 for the intensity 0.10. The conclusion is that the predictive value did not change much and that the interpretation of the results would not be dramatically changed.

6. Application of the multivariate OutbreakP method to Swedish regional influenza data

There are several national and international institutes that collect data on epidemic diseases, for example, the European Centre for Disease Prevention and Control in Europe and the Centers for Disease Control and Prevention in the USA. Monitoring influenza in Sweden is mostly based on reports from all Swedish laboratories providing laboratory diagnoses of influenza (LDI). We will use these LDI data to illustrate the proposed method. In Sweden, data of infectious diseases are collected by the Swedish Institute for Infectious Disease Control, SMI. Descriptions of the collection of these data are given in [1,2]. Here, we use the laboratory-confirmed incidences of

influenza type A or B. For some purposes, it may be of interest to monitor each location separately. However, the aim here is to get an alarm when the influenza epidemic has reached any part of Sweden. This means that the aim is to detect the first outbreak.

6.1 *The spreading pattern of influenza in Sweden*

The spatial pattern of how a disease spreads between regions is important. Spatial clustering of adverse health events is discussed, for example, in [24,27,29,33,38]. However, in some situations, such as in the case of influenza in Sweden, the outbreak pattern is not characterized by clustering.

The spread of epidemic diseases, such as influenza, often follows geographical patterns. Schiöler [35] searched for geographical patterns in the spread of influenza in Sweden (for example, a pattern from South to North or from West to East). No such pattern was found. Instead, it was found that influenza epidemics tend to start in the larger cities and then spread to the smaller ones. Data from areas classified as metropolitan areas generally showed an earlier outbreak than those from the locality areas. The metropolitan areas have major international airports nearby (Arlanda, Landvetter, Umeå and Kastrup), and commuting to other countries is common. This is a plausible explanation for the early start of the influenza season in these areas. This is also in accordance with the results of Crepey and Barthelemy [6], who investigated the relation between traveling and influenza in the USA and in France and found a stable impact.

The time difference in the onset of the influenza outbreak was about 1 week. This information will be used to increase the efficiency of our surveillance system.

6.2 *Outbreak detection of influenza in Sweden*

Based on the results on sufficiency in Section 3, the maximum-likelihood estimation in Section 4.6, the GLR in Section 4.7 and the choice of alarm limit in Section 5 to give $\mathrm{MRL}^0 = 780$ and a predictive value greater than 95%, we applied the OutbreakPSuffR1 to 11 seasons of influenza.

Figure 6 shows the results for the season 2006–2007. By accumulating the information by the OutbreakPSuffR1 alarm statistic, the outbreak is more clearly seen than when by the statistic based on the total number of cases in Sweden.

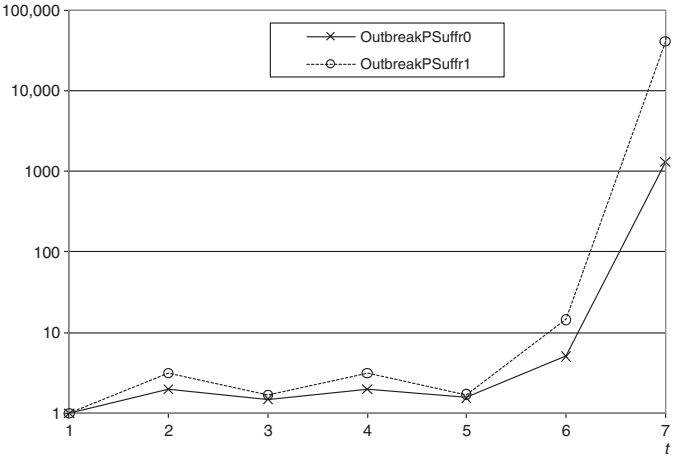

Figure 6. The alarm statistic of the OutbreakPSuffR1 method compared with that of the OutbreakPSuffR0 up to the week of alarm during the season 06–07.

Table 2. Results for 11 influenza seasons in Sweden.

Season	SuffR0	SuffR1	First
1999–2000	49	49	Same
2000–2001	52	52	Same
2001–2002	2	2	Same
2002–2003	1	1	Same
2003–2004	46	46	Same
2004–2005	50	48	SuffR1
2005–2006	1	1	Same
2006–2007	47	46	SuffR1
2007–2008	51	51	Same
2008–2009	48	48	Same
2009–2010	No alarm	24	SuffR1

Notes: The week of alarm is given for the methods based on the SuffR0 and SuffR1, respectively. The last column shows which method gave the first indication of an outbreak.

The situation varies from year to year. In Table 2, the week of the alarm is given for OutbreakP-SuffR0 and OutbreakPSuffR1 for all years with available data. The OutbreakP based on SuffR1 gives an alarm the same week or earlier compared with OutbreakP based on the SuffR0, the total. As can be seen from the table, the alarm is given at the same time for eight seasons and earlier for three seasons for OutbreakP based on SuffR1 when compared with SuffR0. Note that the last season differs from the earlier ones due to the new H1N1 influenza. The incidences (of influenza type A or B) were very low this season and highly dominated by the metropolitan areas. This explains why there was an alarm of an outbreak by the OutbreakSuffR1 method, which utilizes information on the metropolitan areas, but not by OutbreakSuffR0, which uses only the total for the country as a whole.

7. Discussion

In recent years, there have been several events that highlight the importance of outbreak detection. The outbreaks of new kinds of influenza (SARS, avian and H1N1) are such recent examples.

The semi-parametric method used here detects outbreaks defined as a monotonic increase following the constant level before the onset of the outbreak. Such outbreaks are of interest in connection with several diseases and syndromes. Often, the information about the baseline is limited. Errors in the estimation of the baseline can have serious effect, as demonstrated, for example, by Frisén and Andersson [10]. Therefore, it can be of value to have access to a method, which does not require knowledge about the baseline but is focussed on the increasing incidence at an outbreak. A semi-parametric maximum-likelihood ratio surveillance method was derived in [10] for the regular exponential family and applied and compared in [12]. The likelihood principle makes it possible to include knowledge on the probability of an outbreak depending on the season. However, here we chose a non-informative approach, since it may be valuable to detect outbreaks that occur at unexpected times.

When data from different sources are available, multivariate surveillance should be applied. This is the case for detection of influenza outbreaks on the basis of data from different regions. The two simplest approaches of multivariate surveillance are the reduction to a suitable univariate statistic and parallel surveillance with due concern to the multiplicity. We included these approaches in our evaluations by simulations. We also suggested a joint GLR method based on maximum

likelihood under multivariate monotonicity restrictions. The properties depend heavily on the relation between the times of onset in the different processes.

The relation between different processes is important in multivariate surveillance, as demonstrated by, for example, Frisén *et al*. [13]. The method that is optimal for simultaneous changes is not efficient at a time lag. The exact relation between the onset on different location is seldom exactly known. However, there can be some information as demonstrated in, e.g. [35] where it was found that the influenza outbreak in Sweden in general started a week earlier in major cities than the rest of the country. In the application to the Swedish influenza data, it was demonstrated that the performance of the surveillance was improved by utilizing this knowledge. The simulation study demonstrated that even if the true time lag is only approximately known, it was an improvement to use it in the method for the case studied.

Most theory of statistical surveillance is based on a change between two distributions – one for the times before the change point and the other for the times after it. For simultaneous changes, we demonstrated that the sufficient statistic has one change point and that the suggested method is optimal. However, when changes occur at different times, we can have several changes in the multivariate distribution. Thus, we cannot expect optimality. Here, we demonstrated that the suggested method gave good results both in the simulation study and when applied to spatial information on influenza in Sweden. We used a simulation model mimicking the behavior of Swedish influenza data, based on the results of Andersson *et al*. [1], where a discussion on data quality problems was included. When evaluating methods for online monitoring, it is important to use measures that incorporate the time issue, i.e. the fact that there are repeated decisions, not just one decision as in hypothesis testing. Here, we used evaluation measures by Frisén *et al*. [13], which are better suited for multivariate online surveillance than the conventional ones.

The primary motive for this paper was the need for spatial surveillance of influenza outbreaks in Sweden. The suggested method may also be useful for other applications. The case of proxy data for influenza was discussed in Section 2.2. The detection of a change from a constant level to a monotonic trend is of special interest in connection with outbreaks of epidemic diseases. However, it may be useful also in other areas. For example, Schiöler and Frisén [36] discussed the application of the outbreak method for detecting a decline in the results of financial managers.

Acknowledgements

We are grateful for constructive referee comments. Eva Andersson and Kjell Pettersson have given constructive comments. The data were made available to us by the Swedish Institute for Infectious Disease Control, and we are grateful for discussions about the aims and the data quality. The work was supported by the Swedish Civil Contingencies Agency (grant 0314/206).

References

[1] E. Andersson, D. Bock, and M. Frisén, *Modeling influenza incidence for the purpose of on-line monitoring*, Stat. Methods Med. Res. 17 (2008), pp. 421–438.

[2] E. Andersson, S. Kuhlmann-Berenzon, A. Linde, L. Schiöler, S. Rubinova, and M. Frisén, *Predictions by early indicators of the time and height of yearly influenza outbreaks in Sweden*, Scand. J. Public Health 36 (2008), pp. 475–482.

[3] S. Bersimis, S. Psarakis, and J. Panaretos, *Multivariate statistical process control charts: An overview*, Qual. Reliab. Eng. Int. 23 (2007), pp. 517–543.

[4] B. Cakici, K. Hebing, M. Grünewald, P. Saretok, and A. Hulth, *CASE – A framework for computer supported outbreak detection*, BMC Med. Inform. Decis. Mak. 10: 14 (2010).

[5] J.T. Chang and R.D. Fricker, *Detecting when a monotonically increasing mean has crossed a threshold*, J. Qual. Technol. 31 (1999), pp. 217–234.

[6] P. Crepey and M. Barthelemy, *Detecting robust patterns in the spread of epidemics: A case study of influenza in the United States and France*, Amer. J. Epidemiol. 166 (2007), pp. 1244–1251. Available at http://aje.oxfordjournals.org/cgi/content/abstract/166/11/1244.

[7] M. Frisén, *Evaluations of methods for statistical surveillance*, Stat. Med. 11 (1992), pp. 1489–1502.

[8] M. Frisén, *Statistical surveillance. Optimality and methods*, Int. Stat. Rev. 71 (2003), pp. 403–434.

[9] M. Frisén, *Optimal sequential surveillance for finance, public health and other areas. Editor's special invited paper*, Sequential Anal. 28 (2009), pp. 310–337 (discussion 338–393).

[10] M. Frisén and E. Andersson, *Semiparametric surveillance of monotonic changes*, Sequential Anal. 28 (2009), pp. 434–454. Available at http://www.informaworld.com/10.1080/07474940903238029.

[11] M. Frisén, E. Andersson, and K. Pettersson, *Semiparametric estimation of outbreak regression*, Statist. J. Theor. Appl. Statist. 44 (2010), pp. 107–117. Available at http://www.informaworld.com/10.1080/02331880903021484.

[12] M. Frisén, E. Andersson, and L. Schiöler, *Robust outbreak surveillance of epidemics in Sweden*, Stat. Med. 28 (2009), pp. 476–493.

[13] M. Frisén, E. Andersson, and L. Schiöler, *Evaluation of multivariate surveillance*, J. Appl. Stat. 37 (2010), pp. 2089–2100.

[14] M. Frisén, E. Andersson, and L. Schiöler, *Sufficient reduction in multivariate surveillance*, Comm. Statist. Theory Methods 40 (2011), pp. 1821–1838.

[15] M. Frisén and C. Sonesson, *Optimal surveillance based on exponentially weighted moving averages*, Sequential Anal. 25 (2006), pp. 379–403.

[16] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, *Detecting influenza epidemics using search engine query data*, Nature 457 (2009), pp. 1012–1014. Available at http://dx.doi.org/10.1038/nature07634.

[17] L. Held, M. Hofman, M. Höhle, and V. Schmid, *A two-component model for counts of infectious diseases*, Biostatistics 7 (2006), pp. 422–437.

[18] M. Höhle, *Aberration detection in R illustrated by Danish mortality monitoring*, in *Biosurveillance*, T. Kass-Hout and X. Zhang, eds., CRC Press, Boca Raton, FL, 2010, pp. 215–238.

[19] M. Höhle and M. Paul, *Count data regression charts for the monitoring of surveillance time series*, Comput. Statist. Data Anal. 52 (2008), pp. 4357–4368.

[20] A. Hulth, G. Rydevik, and A. Linde, *Web queries as a source for syndromic surveillance*, PLoS ONE 4 (2009), p. e4378. Available at http://dx.doi.org/10.1371%2Fjournal.pone.0004378.

[21] E. Järpe, *On univariate and spatial surveillance*, Ph.D. thesis, Göteborg University, 2000.

[22] M.D. Joner Jr, W.H. Woodall, M.R. Reynolds Jr, and R.D. Fricker, *A one-sided MEWMA chart for health surveillance*, Qual. Reliab. Eng. Int. 24 (2008), pp. 503–518.

[23] T. Kass-Hout and X. Zhang (eds.), *Biosurveillance: A Health Protection Priority*, CRC Press, Boca Raton, FL, 2010.

[24] M. Kulldorff, *Prospective time periodic geographical disease surveillance using a scan statistic*, J. R. Stat. Soc. A 164 (2001), pp. 61–72.

[25] T.L. Lai, *Sequential changepoint detection in quality control and dynamical systems*, J. R. Stat. Soc. B 57 (1995), pp. 613–658.

[26] A.B. Lawson and K. Kleinman (eds.), *Spatial and Syndromic Surveillance for Public Health*, Wiley, New York, NY, 2005.

[27] A. Lawson and C. Rodeiro, *Developements in general and syndromic surveillance for small area health data*, J. Appl. Stat. 31 (2004), pp. 397–406.

[28] C.A. Lowry, W.H. Woodall, C.W. Champ, and S.E. Rigdon, *A multivariate exponentially weighted moving average control chart*, Technometrics 34 (1992), pp. 46–53.

[29] J.B. Marshall, D.J. Spitzner, and W.H. Woodall, *Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time*, Stat. Med. 26 (2007), pp. 1579–1593.

[30] M. Paul, L. Held, and A.M. Toschke, *Multivariate modelling of infectious disease surveillance data*, Stat. Med. 27 (2008), pp. 6250–6267.

[31] T. Robertson, F.T. Wright, and R.L. Dykstra, *Order Restricted Statistical Inference*, Wiley, Chichester, 1988.

[32] P.A. Rogerson, *Surveillance systems for monitoring the development of spatial patterns*, Stat. Med. 16 (1997), pp. 2081–2093.

[33] P.A. Rogerson, *Monitoring point patterns for the development of space–time clusters*, J. R. Stat. Soc. A 164 (2001), pp. 87–96.

[34] T.P. Ryan, *Statistical Methods for Quality Improvement*, 2nd ed., Wiley, New York, NY, 2000.

[35] L. Schiöler, *Characterisation of influenza outbreaks in Sweden*, Scand. J. Public Health 39 (2011), pp. 427–436.

[36] L. Schiöler and M. Frisén, *On statistical surveillance of the performance of fund managers*, Report No. 2008:4, Statistical Research Unit, Department of Economics, University of Gothenburg, Sweden, 2008.

[37] G. Shmueli and H.S. Burkom, *Statistical challenges facing early outbreak detection in biosurveillance*, Technometrics 52 (2010), pp. 39–51.
[38] C. Sonesson, *A CUSUM framework for detection of space–time disease clusters using scan statistics*, Stat. Med. 26 (2007), pp. 4770–4789.
[39] C. Sonesson and D. Bock, *A review and discussion of prospective statistical surveillance in public health*, J. R. Stat. Soc. A 166 (2003), pp. 5–21.
[40] C. Sonesson and M. Frisén, *Multivariate surveillance*, in *Spatial Surveillance for Public Health*, A. Lawson and K. Kleinman, eds., Wiley, New York, NY, 2005, pp. 169–186.
[41] P. Wessman, *Some principles for surveillance adopted for multivariate processes with a common change point*, Comm. Statist. Theory Methods 27 (1998), pp. 1143–1161.
[42] W.H. Woodall, *The use of control charts in health-care monitoring and public-health surveillance*, J. Qual. Technol. 38 (2006), pp. 89–134.
[43] W.H. Woodall and D.C. Montgomery, *Research issues and ideas in statistical process control*, J. Qual. Technol. 31 (1999), pp. 376–386.
[44] H. Zhou and A.B. Lawson, *EWMA smoothing and Bayesian spatial modeling for health surveillance*, Stat. Med. 27 (2008), pp. 5907–5928.
[45] Q. Zhou, Y. Luo, and Z. Wang, *A control chart based on likelihood ratio test for detecting patterned mean and variance shifts*. Comput. Statist. Data Anal. 54 (2010), pp. 1634–1645.

Appendix 1. Proof of Theorem 1

Since the observations are independent given the values of the change points, the density can be written as a product. We will first consider a fixed but unknown value of τ_1. The likelihood expressions for the one-parameter exponential family can be written as

$$f(Y;\tau_1 \le s) = \exp\left\{\sum_{i=1}^{p}\sum_{t=1}^{\min(\tau_i-1,s)}[y_i(t)(\varphi_0)+g(\varphi_0)+h(y_i(t))] + \sum_{i=1}^{p}\sum_{t=\tau_i}^{s}[y_i(t)(\varphi_{t-\tau_i+1})+g(\varphi_{t-\tau_i+1})+h(y_i(t))]\right\}$$

and

$$f(Y;\tau_1 > s) = \exp\left\{\sum_{t=1}^{s}\sum_{j=1}^{p}[y_j(t)(\varphi_0)+g(\varphi_0)+h(y_j(t))]\right\}.$$

Thus, we have the partial log-likelihood ratio

$$\begin{aligned}L(s,\tau_1) = \log\frac{f(Y;\tau_1\le s)}{f(Y;\tau_1>s)} &= \sum_{i=1}^{p}\sum_{t=1}^{\min(\tau_i-1,s)}[y_i(t)(\varphi_0)+g(\varphi_0)+h(y_i(t))]\\ &+\sum_{i=1}^{p}\sum_{t=\tau_i}^{s}[y_i(t)(\varphi_{t-\tau_i+1})+g(\varphi_{t-\tau_i+1})+h(y_i(t))]\\ &-\sum_{t=1}^{s}\sum_{i=1}^{p}[y_i(t)(\varphi_0)+g(\varphi_0)+h(y_i(t))]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{p} \sum_{t=\tau_i}^{s} [y_i(t)(\varphi_{t-\tau_i+1}) - y_i(t)(\varphi_0) + g(\varphi_{t-\tau_i+1}) - g(\varphi_0)] \\
&= \sum_{i=1}^{p} \sum_{t=\tau_1+q_i}^{s} [y_i(t)(\varphi_{t-(\tau_1+q_i)+1} - \varphi_0)] + z(\varphi_0, \ldots, \varphi_{s-\tau_1+1}) \\
&= \sum_{i=1}^{p} \sum_{t=\tau_1}^{s-q_i} [y_i(t+q_i)(\varphi_{t-\tau_1+1} - \varphi_0)] + z(\varphi_0, \ldots, \varphi_{s-\tau_1+1}) \\
&= \sum_{t=\tau_1}^{s} \sum_{i \in I_t} [y_i(t+q_i)(\varphi_{t-\tau_1+1} - \varphi_0)] + z(\varphi_0, \ldots, \varphi_{s-\tau_1+1}) \\
&= \sum_{t=\tau_1}^{s} (\varphi_{t-\tau_1+1} - \varphi_0) \sum_{i \in I_t} [y_i(t+q_i)] + z(\varphi_0, \ldots, \varphi_{s-\tau_1+1}),
\end{aligned}
$$

which depends on the observations only through the statistic in the theorem. Since the likelihood ratio is sufficient for the problem, by Halmos factorization criterion, the statistic is also sufficient. This completes the proof when τ_1 is fixed but unknown.

If τ_1 is stochastic with some density g, then the partial log-likelihood ratio can be written as

$$L(s) = \sum_{r=1}^{\infty} g(r) L(s,r) = \sum_{r=1}^{s} g(r) L(s,r) + \sum_{r=s+1}^{\infty} g(r) L(s,r),$$

where for each $\tau_1 > s$ no change occur and the partial-likelihood ratio is one and thus $L(s,r) = 0$. Thus,

$$L(s) = \sum_{r=1}^{s} g(r) \left[\sum_{t=r}^{s} (\varphi_{t-r+1} - \varphi_0) \sum_{i \in I_t} [y_i(t+q_i)] \right] + z^*(\varphi_0, \ldots, \varphi_s) + 0,$$

and again we have, by the factorization theorem, that the statistic in Theorem 1 is sufficient for the problem.

Appendix 2. Proof of Theorem 2

In order to obtain the maximum-likelihood estimators of the expected values λ_t for $\tau_1 = 1$, we utilize the assumption $\lambda_0 \leq \lambda_1 \cdots \leq \lambda_s$. In [11], it was demonstrated that in the univariate case, the maximum-likelihood estimators of the expected values λ_t of the outbreak regression can be obtained by the PAVA algorithm. For p processes, with known lags $(q_1 = 0, q_2, q_3, \ldots, q_p)$, any observation of $Y_i(t)$ such that $t < \tau_i$ is an observation with the expected value λ_0. In the same way, any observation of $Y_i(t)$ such that $\tau_i = t$ has the expected value λ_1 and so on until the last observations of $Y_1(s)$ and any other $Y_i(s)$ such that $\tau_i = \tau_1$, which are observations with the expected value λ_s. Thus, the number of processes observed, $|I_t|$, with expectation λ_t depends on t and $(q_2, q_3, \ldots, q_p)$.

Now we use results on isotonic regression, with different numbers of observations for different values of the independent variable. We reformulate Theorem 1.5.2 in [26] for the one-parameter case and our parameterization. It states that, assuming independent random samples from each of s

populations characterized by regular exponential densities of the form $f(y_i(t);\varphi_t) = y_i(t)(\varphi_t) + g(\varphi_t) + h(y_i(t))$, the maximum-likelihood estimates under the order restriction $Y_i(1) \leq \cdots \leq Y_i(s)$ is given by applying the PAVA algorithm to $n_j^{-1}\sum_{j=1}^{n_j} Y_i(t)$ with weights n_j, where n_j denotes the sample size of population j. We utilize this theorem by observing that by our definition we at each time t observe $|I_t|$ iid processes (i.e. a sample of $|I_t|$ observations) characterized by the exponential density above with parameter φ_t. We observe the processes at s different time points (i.e. we have s populations) and hence we obtain the maximum-likelihood estimates under the order restriction $Y_i(1) \leq \cdots \leq Y_i(s)$ by the PAVA algorithm.