


Research Report
Statistical Research Unit
Department of Economics
University of Gothenburg
Sweden

Multivariate outbreak detection

Schiöler, L. and Frisén, M.

Research Report
2010:2
ISSN 0349-8034

Mailing address:
Statistical Research Unit
P.O. Box 640
SE 405 30 Göteborg
Sweden

Fax
Nat: 031-786 12 74
Int: +46 31 786 12 74

Phone
Nat: 031-786 00 00
Int: +46 31 786 00 00

Home Page:
http://www.statistics.gu.se/

Multivariate outbreak detection

BY LINUS SCHIÖLER[1] and MARIANNE FRISÉN[1]

University of Gothenburg

On-line monitoring is needed to detect outbreaks of diseases like 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 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.

1. Introduction

On-line 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 by Lai (1995), Woodall and Montgomery (1999), Ryan (2000), Frisén (2003), Frisén (2009) 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 on-line surveillance. Surveillance in public health is reviewed by for example Sonesson and Bock (2003), Lawson and Kleinman (2005), Woodall (2006), Shmueli and Burkom (2010), and Kass-Hout and Zhang (2010). 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 reasons very costly to society and it is therefore of great value to monitor the epidemic period in order to properly allocate medical resources (Andersson et al. (2008b)). A semi-parametric method for detecting the onset of a monotonic increase was suggested for univariate surveillance by Frisén and Andersson (2009). It was successfully applied to the incidence of influenza in Sweden as a whole by Frisén et al. (2009).

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 by Sonesson and Frisén (2005) and Joner Jr. et al. (2008). 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 by Schiöler (2010) 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

[1] Supported by the Swedish Civil Contingencies Agency (grant 0314/206).

Key words and phrases. Exponential family, Generalised likelihood, Ordered regression, Spatial data, Surveillance.

outbreak surveillance system. Another case there 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 Hulth et al. (2009) and Ginsberg et al. (2009) it was shown that data of search patterns on the Internet could be used as a proxy for influenza incidence. Ginsberg, et al. (2009) found that the lag in reporting was about one day compared to between one and two weeks for traditional CDC-data. The method suggested in this article 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 the various locations.

In Section 2, we will specify univariate and multivariate models for outbreaks. In Section 3, we will derive a sufficient reduction of the data for multivariate outbreak situations. Sufficient reduction for detection of step changes was earlier derived by Frisén et al. (2010c) but here it is derived for 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 by Frisén and Andersson (2009). 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 method. In Section 5, we evaluate the suggested method by a simulation study, where properties like 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. (2010b) 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 the final section.

2. Specification of the outbreak model

At each time point, t, a new observation is made on a process **Y**. 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 Andersson et al. (2008a) 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 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 $\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_s$. 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.

The situation where the regression is constant at first and then monotonically increasing will be called “outbreak regression”.

2.2. *Multivariate outbreak*

In multivariate surveillance the process under surveillance is a p-variate vector, denoted by $\mathbf{Y} = \{\mathbf{Y}(t), \quad t = 1,2,...\}$, where $\mathbf{Y}(t) = \{Y_1(t), Y_2(t),..., Y_p(t)\}$. The components of the vector represent, for example, the incidence of a disease at p different locations. Each component has the same properties as $\lambda(t)$ 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, $\mathbf{Y}^s = \{\mathbf{Y}(1), \mathbf{Y}(2)... \mathbf{Y}(s)\}$.

When several processes are observed, knowledge about the relation between the times of the onsets of the outbreaks is essential. Different methods are suitable for different relations. The aim is to detect an outbreak in any of the processes, which means that we aim at detecting the first one. The time τ_i of the onset of the outbreak of process Y_i may not be the same for all i=1,...p. The relation between the times is important. 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 misspecified is examined in Section 5.5. For notational convenience we order the processes according to which changes first, so that $\tau_1 \leq ... \leq \tau_p$, and denote the time lag for process Y_i by q_i, where $q_1=0$ and $q_i = \tau_i - \tau_1$ for i=2,...,p. The case where the onsets are simultaneous, that is $\tau_i = \tau$ for i=1,...p, is of special interest. In this case $q_i=0$ i=1,...p. We denote this by lag=0. In numerical examples and applications we will also use the special cases of two processes with $q_2=1$ or $q_2=2$. We denote this by lag=1 and lag=2, respectively.

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.

If a parametric shape of the outbreak pattern is known, this should be used to increase efficiency. However, 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.

3. Sufficient reduction at multivariate outbreaks

In Frisén, et al. (2010b) it was demonstrated that the relation between the change points of the different processes is very important, since it affects the properties of different surveillance methods in different ways. In simple examples, it was demonstrated that a method which is optimal for simultaneous changes is inefficient in other cases. Thus, any knowledge on the change points should be utilized. A sufficient reduction will not reduce the information and still allows a joint solution to the full surveillance problem. It is of special interest to study a simultaneous outbreak at all locations and also a time lag in the onset of the outbreaks. 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 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.

Wessman (1998) and Frisén, et al. (2010c) demonstrated that if all processes have the same change points, i.e. $\tau_1= \tau_2=...\tau_p=\tau$, then the univariate vector of partial likelihood ratios, {L(s,t), t=1,...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. Zhou et al. (2010) used this result 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 Frisén and Andersson (2009) and Frisén, et al. (2009).

3.2. *Changes with a time lag between locations*

Järpe (2000) studied the case of a known time lag for independent normal distributions with equally sized shifts in the expected value at the change points and demonstrated that a sufficient reduction to univariate surveillance exists. Frisén, et al. (2010c) studied the case of changes in the general one-parameter exponential family (including the Poisson distribution) but also only for step changes. Different levels of the parameter before the change as well as differences in shift size were considered.

The earlier results on sufficiency for the detection of a step change cannot be used directly for outbreak detection, since we are interested in detecting a change from a constant level to a monotonically increasing one rather than a sudden shift. Here, we study the case 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 $\mathbf{Y}_1$. Theorem 1 shows that a sufficient reduction to a univariate statistic exists for the situation with different outbreak times, and in Example 1 (after Theorem 1 and its proof) the theorem is illustrated for a simple case. 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 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

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

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

Proof: Since the observations are independent given the values of the change points, the distribution 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

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

and

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

Thus, we have the log likelihood ratio

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

$$+\sum_{i=1}^{p}\sum_{t=\tau_i}^{s}\left[y_i(t)(\varphi_{t-\tau_i+1}) + g(\varphi_{t-\tau_i+1}) + h(y_i(t))\right] - \sum_{t=1}^{s}\sum_{i=1}^{p}\left[y_i(t)(\varphi_0) + g(\varphi_0) + h(y_i(t))\right]$$

$$= \sum_{i=1}^{p}\sum_{t=\tau_i}^{s}\left[y_i(t)(\varphi_{t-\tau_i+1}) - y_i(t)(\varphi_0) + g(\varphi_{t-\tau_i+1}) - g(\varphi_0)\right]$$

$$= \sum_{i=1}^{p}\sum_{t=\tau_1+q_i}^{s}\left[y_i(t)(\varphi_{t-(\tau_1+q_i)+1} - \varphi_0)\right] + z(\varphi_0,...,\varphi_{s-\tau_1+1})$$

$$= \sum_{i=1}^{p}\sum_{t=\tau_1}^{s-q_i}\left[y_i(t+q_i)(\varphi_{t-\tau_1+1} - \varphi_0)\right] + z(\varphi_0,...,\varphi_{s-\tau_1+1})$$

$$= \sum_{t=\tau_1}^{s}\sum_{i\in I_t}\left[y_i(t+q_i)(\varphi_{t-\tau_1+1} - \varphi_0)\right] + z(\varphi_0,...,\varphi_{s-\tau_1+1})$$

$$= \sum_{t=\tau_1}^{s}(\varphi_{t-\tau_1+1} - \varphi_0)\sum_{i\in I_t}\left[y_i(t+q_i)\right] + z(\varphi_0,...,\varphi_{s-\tau_1+1}),$$

which depends on the observations only through the statistic in the theorem. The likelihood ratio is sufficient for the problem, and hence the statistic is sufficient. This completes the proof when τ_1 is fixed but unknown.

If τ_1 is stochastic with some distribution g(t), then the density of Y can be written:

$$f(Y) = \sum_{t=1}^{\infty} g(t)\, f(Y \mid \tau_1 = t),$$

which is a function of $f(Y \mid \tau_1 = t)$, and hence the arguments above can be used to show that the statistic in Theorem 1 is sufficient for the problem also in this case.■

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

$$\sum_{i\in I_t} Y_i(t+q_i)/\mid I_t \mid : t = 1,...,s,$$

where $\mid I_t \mid$ denotes the cardinality of I_t, is also 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, SuffRq(s,t)= $\sum_{i\in I_t} Y_i(t+q_i)/\mid I_t \mid : t = 1,...,s$,where q is the lag between the two processes.

EXAMPLE 1. For two processes Y1 and Y2 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 {$Y_1(1)+Y_2(2)$, $Y_1(2)+Y_2(3)$,...., $Y_1(s-1)+Y_2(s)$, $Y_1(s)$ } at s. A numerical example is given in Section 4.6. ■

The sufficient statistic at decision time s is SuffRq(s,t) t=1,...s, where SuffRq(s,t)=$\left(Y_1(t)+Y_2(t+q)\right)/2$ for t≤s -q and SuffRq(s,t)=$Y_1(t)$ for t>s-q. In Example 1 we have {SuffR1(1,t)}= {{Y1(1)} at s=1. At s=2 we have {SuffR1(2,t)}= $\{[Y_1(1)+Y_2(2)]/2, Y_2(2)\}$. At s=3 we have {SuffR1(3,t)}= {{Y1(1)+Y2(2)]/2, [Y1(2)+Y2(3)]/2, Y1(3)}. More generally we have {SuffRp1(p,t)}= {[$Y_1(1)+Y_2(2)$]/2, ...[$Y_1(2)+Y_2(3)$]/2,...., [$Y_1(s-1)+Y_2(s)$]/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 (2009). 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 generalized likelihood ratio 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 for example Held et al. (2006). 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.

For the case of unknown parameters, generalized likelihood ratios (GLR) can be used by substituting the parameters with the maximum likelihood estimates. Lai (1995) 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 (2008) for Poisson and negative binomial distribution at surveillance of infectious diseases. In Frisén and Andersson (2009) a 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 Frisén et al. (2010a) for the 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. The method is semi-parametric since the distribution is parametric, but the regression is non-parametric since the only restriction on the regression 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 Höhle (2010) and available on CRAN, and the open `JAVA package CASE` described in Cakici et al. (2010).

For the univariate surveillance of the influenza incidence in Sweden as a whole, the OutbreakP method was evaluated by Frisén and Andersson (2009) and Frisén, et al. (2009). 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. (1992). When the multivariate distribution is available, as in e.g. Paul (2008), this might be used as a base for a surveillance method. An important situation treated by e.g. Tartakovsky and Veeravalli (2008) is where change in only one location can be expected and the identification of the correct one is crucial. General reviews on multivariate surveillance methods can be found for example in Basseville and Nikiforov (1993), Sonesson and Frisén (2005), Bersimis et al. (2007) and Frisén (2010).

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. In Sullivan and Jones (2002) this is referred to as "scalar accumulation". One natural reduction when dealing with multivariate normal variables is to use the Hotelling T^2 statistic suggested by Hotelling (1947). The Hotelling T^2 statistic is defined as $T^2(t) = (\mathbf{Y}(t) - \boldsymbol{\mu}_0(t))^T \mathbf{S}_{\mathbf{Y}(t)}^{-1} (\mathbf{Y}(t) - \boldsymbol{\mu}_0(t))$, where $\mathbf{S}_{\mathbf{Y}(t)}$ is the sample covariance matrix. Originally, the Hotelling T^2 statistic was used in a Shewhart approach, and this is sometimes referred to as the Hotelling T^2 control chart.

One example of scalar accumulation is when, for each time point, a statistic representing the important aspects of the spatial pattern is constructed from a purely spatial analysis. 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 Rogerson (1997) and Rogerson (2001), different statistics measuring clustering were used for each time, and the information was accumulated by the univariate CUSUM method. In Zhou and Lawson (2008), 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 natural 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 Frisén and Andersson (2009). 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. By the parallel approach, 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. An overall alarm is called the first time that any of the processes gives an

alarm. The method is called OutbreakPParallel. Results for this method, as compared to others, are given in Section 5.3.

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

The likelihood ratio of an outbreak versus no outbreak with onsets of the outbreaks at $\tau_1, \tau_2, \ldots \tau_p$, is

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

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

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

For detection of an outbreak as defined in Section 2 L(s,1) is the relevant statistic, see Frisén and Andersson (2009). 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 $I_t = \{i : q_i \le s - t, 1 \le i \le p\}$.

For two processes we have

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

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 generalized likelihood ratio 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

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

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. (1988).

Theorem 2: For the multivariate outbreak regression in Section 2.2 with processes which all belong to the one-parameter 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), 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.

Proof:
In order to obtain the maximum likelihood estimators of the expected values λ_t for $\tau_1=1$, we utilize the assumption $\lambda_0 \leq \lambda_1 \ldots \leq \lambda_s$. Frisén, et al. (2010a) 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 observations, $|I_t|$, with expectation λ_t depends on t and $(q_2, q_3, \ldots q_p)$. It follows from results on isotonic regression, with different numbers of observations for different values of the independent variable (see for example Theorem 1.5.2 in Robertson, et al. (1988)), that the maximum likelihood estimators are obtained by the PAVA on the average of the observations of λ_t with weights proportional to the number of observations, $|I_t|$. ■

EXAMPLE 2
To illustrate how the sufficient reduction and PAVA are used, we give a simple example for two processes with lag q=1. SuffRq(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.

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	

The estimate of $\hat{\lambda}_0$ is the average of all observations. At s=5 we have $\hat{\lambda}_0 = 2.6$. To estimate $\hat{\lambda}_t$ at time s=5, we apply the PAVA to the sequence SuffR1(5,t), t=1,...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 obs ervation, $Y_2(1)$, 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. *Generalized likelihood ratio surveillance of multivariate outbreaks*

We will use the generalized likelihood ratio, 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

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

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

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

In the case q=0 this simplifies to the univariate OutbreakP statistic described in Frisén and Andersson (2009) and Frisén, et al. (2009).

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

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

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

In a multivariate situation, some reduction of 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 $t\ge\tau_i$. 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*

We used a model that is relevant for the application to the influenza data described in Section 6. The model is based on the study by Andersson, et al. (2008a) on the seasonal influenza in Sweden. The Poisson distribution was used for the incidences. The suggested method is non-parametric with respect to the shape. However, to examine the properties of the method by a simulation study, we used a parametric model to generate data. For the total influenza incidence in Sweden, the level at the constant phase, λ_0, is set to $\lambda_0 = 1$, and the parameter $\lambda(t)$ of the Poisson distribution follows an exponential curve $\lambda(t) = \exp(\beta_0 + \beta_1(t-\tau+1)$ for the increasing phase. The parameters were estimated to $\beta_0 = -0.26$ and $\beta_1 = 0.826$ from Swedish influenza data from the season 03-04, which was not extreme in any sense but "typical".

For the multivariate case, we use a model with two processes resembling those of the influenza data in Section 6. We use the results by Schiöler (2010) on how the incidence develops for the Metropolitan, M, and Local, L, areas, respectively. We use $E[M(t)]=0.5$ for $t<\tau$ and $E[M(t)]= \exp\{\beta_0+\beta_1(t-\tau+1)\}$, and $E(L(t)= 0.5$ for $t<\tau$ and $E(L(t))= \exp\{\beta_0+\beta_1 (t-\tau+1+q)\}$). With parameters, $\beta_0 = -0.622$ and $\beta_1 = 0.826$.

5.2. *False alarms*

The most commonly used measure for false alarms is the in-control average run length, ARL^0, $E[t_A|\tau=\infty]$. This can be used also in a multivariate situation. A similar measure, which is more convenient to calculate, is the median run length, MRL^0. We used the same MRL^0 (780) in all comparisons in this paper. It was used also for the univariate OutbreakP method in Frisén and Andersson (2009). The technique chosen by Frisén and Sonesson (2006) 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 $E[t_A| \tau_1= \tau_{2=\ldots} \tau_p =1]$. 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.

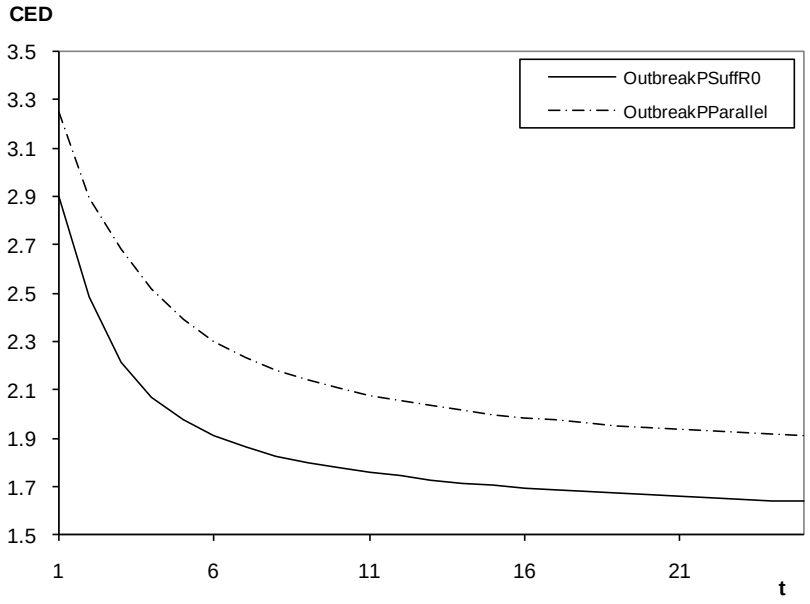

Fig. 1 The conditional expected delay for the OutbreakPParallel and OutbreakPSuffR0 methods for two processes with simultaneous onset of the outbreak (lag=0) as a function of $\tau_{min}=t$.

The conditional expected delay $CED(\tau) = E[t_A - \tau \mid \tau \le t_A]$ can be generalized for multivariate surveillance to CED($\tau_1, \tau_2, \ldots \tau_p$) = $E[t_A - \tau_{\min} \mid \tau_{\min} \le t_A]$, see Frisén, et al. (2010b). For a given lag this depends on only one of the change points. Thus we can write $CED(\tau_{\min}) = E[t_A - \tau_{\min} \mid \tau_{\min} \le t_A]$. 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 we can see that the delay for the parallel method is worse than that for the OutbreakPSuffR1 method based on SuffR1 when lag=1.

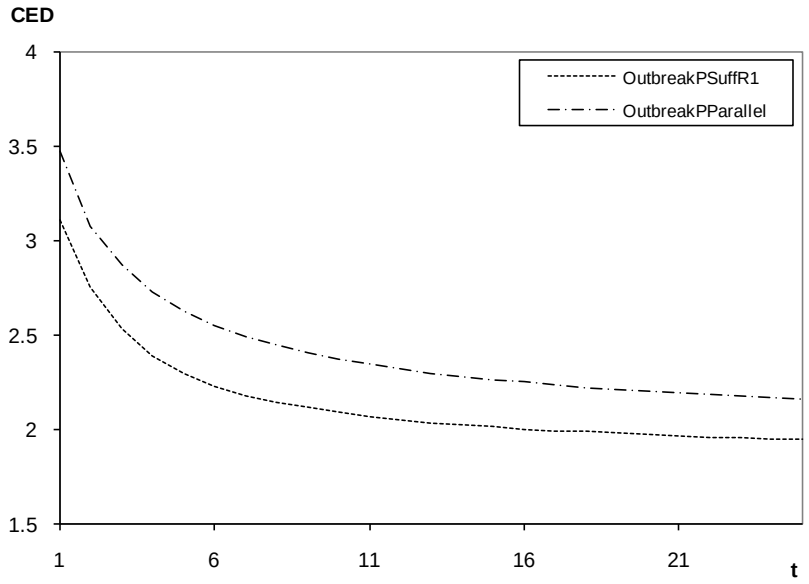

Fig. 2 The delay in detection of the outbreak for the OutbreakPParallel and OutbreakPSuffR1 methods for two processes with lag=1 as a function of τ_{min}=t.

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, as compared to 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. If we have one change point τ and this is regarded as a random variable, this can be done by the probability of an outbreak, at an alarm, as suggested by Frisén (1992):

$$PV(t) = P(\tau \le t \mid t_A = t) = \frac{\sum_{i=1}^{t} P(t_A = t \mid \tau = i) P(\tau = i)}{\sum_{i=1}^{t} P(t_A = t \mid \tau = i) P(\tau = i) + P(t_A = t \mid \tau > t) P(\tau > t)}.$$

In a multivariate setting this was generalized by Frisén, et al. (2010b) to

$$PV(t) = P(\tau_{\min} \le t \,|\, t_A = t) = \frac{\sum_{i=1}^{t}\left(P(t_A = t|\tau_{\min} = i)P(\tau_{\min} = i)\right)}{\sum_{i=1}^{t}\left(P(t_A = t|\tau_{\min} = i)P(\tau_{\min} = i)\right) + P(t_A = t|\tau_{\min} > t)P(\tau_{\min} > t)}.$$

The predictive value depends on whether outbreaks appear frequently or rarely. Knowledge of the exact distribution of τ_{min} is seldom available, but we will nevertheless try to give a rough indicator. In the simulation study, τ_{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 $\nu = 0.1$. 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*

Some models and assumptions are needed in order to efficiently make inferences from data. Hence, it is important to chose 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. We examined the effect of using the sufficient statistic for lag=1 when in fact lag=2, and vice versa. In Figure 3, we have simulated influenza outbreaks where the true lag is 1. We can see that when we used the method

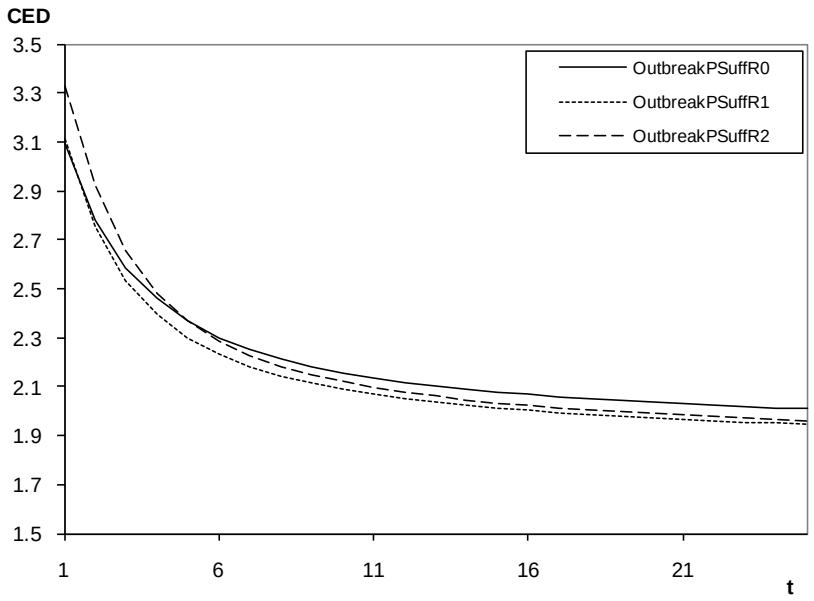

Fig. 3 The delay, as a function of τ_{min}=t, for outbreak detection by OutbreakPSuffR0, OutbreakPSuffR1 and OutbreakPSuffR2 when the true lag is 1.

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. In Figure 4, 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 τ=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, while the true lag was 2, we got clearly worse results. The conclusion is that an approximate lag may work well, provided that it is not too far away from the true one.

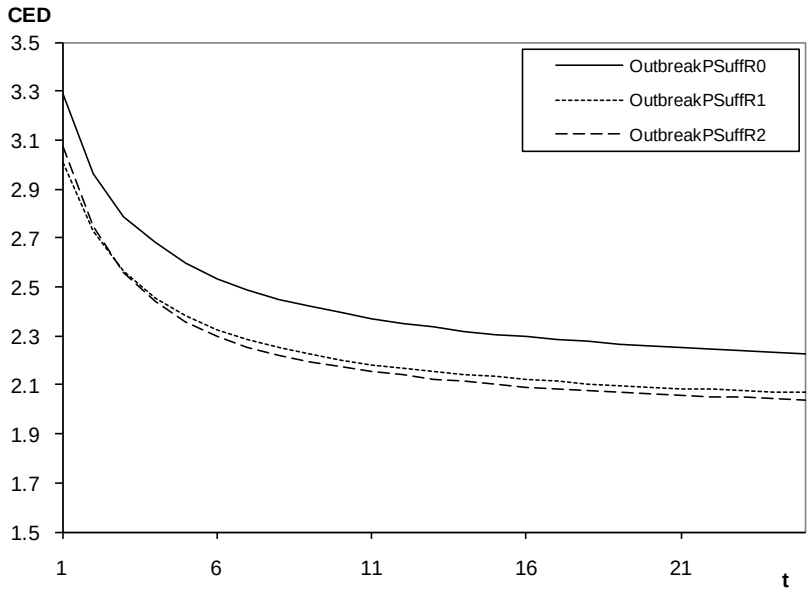

Fig. 4 The delay, as a function of τ_{min}=t, for outbreak detection by OutbreakPSuffR0, OutbreakPSuffR1 and OutbreakPSuffR2 when the true lag is 2.

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 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 Schiöler (2010). We examined what would happen if no adjustments were made and the same weights and 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 US. The monitoring of 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. Andersson, et al. (2008a) and Andersson, et al. (2008b) give descriptions of the collection of these data. 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 by Kulldorff (2001), Rogerson (2001), Lawson and Rodeiro (2004), Marshall et al. (2007) and Sonesson (2007). 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 (2010) 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 (2007), who investigated the relation between travelling and influenza in the US and in France and found a stable impact.

The time difference in the onset of the influenza outbreak was about one 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 generalized likelihood ratio in Section 4.7 and the choice of alarm limit in Section 5 to give $MRL^0=780$ and a predictive value greater than 90 %, we applied the OutbreakPSuffR1 to 11 seasons of influenza.

Figure 5 shows the results for the season 06-07. 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.

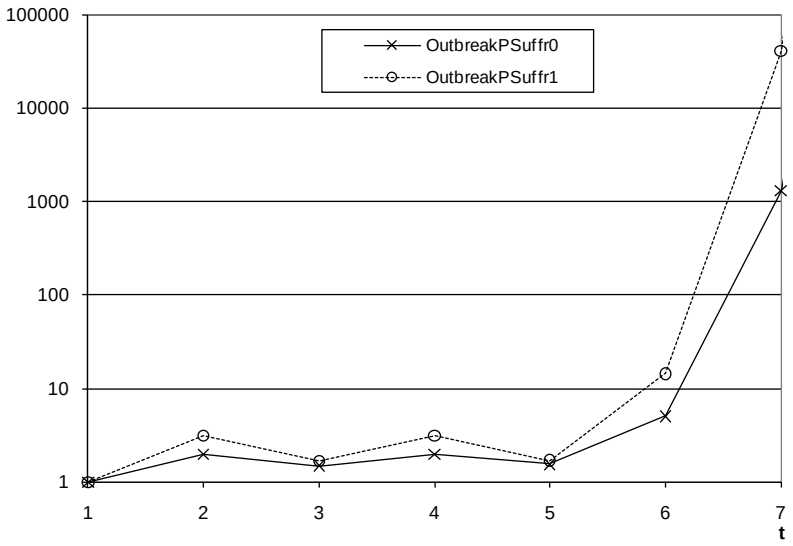

Fig. 5 The alarm statistic of the OutbreakPSuffR1method compared to that of OutbreakPSuffR0 up to the week of alarm during the season 06-07.

The situation varies from year to year. In Table 2, the week of the alarm is given for OutbreakPSuffR0 and OutbreakPSuffR1 for all years with available data. The alarm limits were chosen by way of the simulation study in Section 5 to have the same false alarm property with MRL^0=780. The OutbreakP based on SuffR1 gives an alarm the same week or earlier compared to 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 as compared to 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.

Table 2. Results for 11 influenza seasons in Sweden. 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.

Season	SuffR0	SuffR1	First
99_00	49	49	Same
00_01	52	52	Same
01_02	2	2	Same
02_03	1	1	Same
03_04	46	46	Same
04_05	50	48	SuffR1
05_06	1	1	Same
06_07	47	46	SuffR1
07_08	51	51	Same
08_09	48	48	Same
09_10	No alarm	24	SuffR1

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.

Several different definitions of an outbreak are used, explicitly or implicitly, in literature. Three commonly used approaches to outbreak detection are: i) the detection of an increasing incidence, ii) the detection of an incidence that is higher than expected, based on the information available up to that point and iii) the detection of a spatial clustering of cases which results in a higher incidence in an area than in its surroundings. The choice of method and evaluation procedure depends on which definition is used. Therefore, it is important to state the aim explicitly. Different methods may be optimal under different conditions, which means that the methods can often be seen as complements to each other.

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 (2009). Also, there may be seasonal effects with the same periodicity as the disease as well as large variation between years, thus making it hard to state the expected incidence. Therefore, it can be of value to have access to a method, which does not require knowledge about the baseline but is focused on the increasing incidence at an outbreak. A semi-parametric maximum likelihood ratio surveillance method was derived in Frisén and Andersson (2009) for the regular exponential family and applied and compared in Frisén, et al. (2009). 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 generalized likelihood ratio 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 e.g. Frisén, et al. (2010b). 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. Schiöler (2010) 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 the even if the true time lag is only approximately known it can be an improvement to use it in the method.

Most theory of statistical surveillance is based on a change between two distributions – one for the times before the change point and another 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. (2008a), where a discussion on data quality problems was included. When evaluating methods for on-line 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. (2010b), which are better suited for multivariate on-line 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 (2008) discussed the application of the outbreak method for detecting a decline in the results of financial managers.

Acknowledgements. 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

ANDERSSON, E., BOCK, D. and FRISÉN, M. (2008a). Modeling influenza incidence for the purpose of on-line monitoring. *Statistical Methods in Medical Research,* **17** 421-438.

ANDERSSON, E., KUHLMANN-BERENZON, S., LINDE, A., SCHIÖLER, L., RUBINOVA, S. and FRISÉN, M. (2008b). Predictions by early indicators of the time and height of yearly influenza outbreaks in Sweden. *Scandinavian Journal of Public Health,* **36** 475-482.

BASSEVILLE, M. and NIKIFOROV, I. (1993). *Detection of abrupt changes- Theory and application.* Prentice Hall, Englewood Cliffs.

BERSIMIS, S., PSARAKIS, S. and PANARETOS, J. (2007). Multivariate Statistical Process Control Charts: An Overview. *Quality and Reliability Engineering International,* **23** 517-543.

CAKICI, B., HEBING, K., GRÜNEWALD, M., SARETOK, P. and HULTH, A. (2010). CASE –a framework for computer supported outbreak detection. *BMC Medical Informatics and Decision Making,* **10**.

CREPEY, P. and BARTHELEMY, M. (2007). Detecting Robust Patterns in the Spread of Epidemics: A Case Study of Influenza in the United States and France. *American Journal of Epidemiology,* **166** 1244-1251.

FRISÉN, M. (1992). Evaluations of methods for statistical surveillance. *Statistics in Medicine,* **11** 1489-1502.

FRISÉN, M. (2003). Statistical surveillance. Optimality and methods. *International Statistical Review,* **71** 403-434.

FRISÉN, M. (2009). Optimal sequential surveillance for finance, public health and other areas. Editor's special invited paper. *Sequential Analysis,* **28** 310-337, discussion 338-393.

FRISÉN, M. (2010). Principles for Multivariate Surveillance. In *Frontiers in Statistical Quality Control 9 (*H.-J. LENZ, P.-T. WILRICH and W. SCHMID, eds.) 133-144. Physica-Verlag, Heidelberg.

FRISÉN, M. and ANDERSSON, E. (2009). Semiparametric surveillance of monotonic changes. *Sequential Analysis,* **28** 434-454.

FRISÉN, M., ANDERSSON, E. and PETTERSSON, K. (2010a). Semiparametric estimation of outbreak regression. *Statistics: A Journal of Theoretical and Applied Statistics,* **44** 107 - 117.

FRISÉN, M., ANDERSSON, E. and SCHIÖLER, L. (2009). Robust outbreak surveillance of epidemics in Sweden. *Statistics in Medicine,* **28** 476-493.

FRISÉN, M., ANDERSSON, E. and SCHIÖLER, L. (2010b). Evaluation of Multivariate Surveillance. *Journal of Applied Statistics* to appear.

FRISÉN, M., ANDERSSON, E. and SCHIÖLER, L. (2010c). Sufficient reduction in multivariate surveillance. *Communications in Statistics-Theory and Methods* to appear.

FRISÉN, M. and SONESSON, C. (2006). Optimal surveillance based on exponentially weighted moving averages. *Sequential Analysis,* **25** 379-403.

GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S. and BRILLIANT, L. (2009). Detecting influenza epidemics using search engine query data. *Nature,* **457** 1012-1014.

HELD, L., HOFMAN, M., HÖHLE, M. and SCHMID, V. (2006). A two-component model for counts of infectious diseases. *Biostatistics,* **7** 422-437.

HOTELLING, H. (1947). Multivariate Quality Control. In *Techniques of statistical analysis (*C. EISENHART, M. W. HASTAY and W. A. WALLIS, eds.) 111-184. McGraw-Hill, New York.

HULTH, A., RYDEVIK, G. and LINDE, A. (2009). Web Queries as a Source for Syndromic Surveillance. *PLoS ONE,* **4** e4378.

HÖHLE, M. (2010). Aberration Detection in R Illustrated by Danish Mortality Monitoring. In *Biosurveillance (*T. KASS-HOUT and X. ZHANG, eds.) CRC Press.

HÖHLE, M. and PAUL, M. (2008). Count data regression charts for the monitoring of surveillance time series. *Computational Statistics & Data Analysis,* **52** 4357-4368.

JONER JR., M. D., WOODALL, W. H., REYNOLDS JR., M. R. and FRICKER, R. D. (2008). A One-sided MEWMA Chart for Health Surveillance. *Quality and Reliability Engineering International,* **24** 503-518.

JÄRPE, E. (2000). *On univariate and spatial surveillance. Ph.D Thesis.* Göteborg University, Göteborg.

KASS-HOUT, T. and ZHANG, X. (Eds.). (2010). *Biosurveillance: A Health Protection Priority*: CRC Press.

KULLDORFF, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A,* **164** 61-72.

LAI, T. L. (1995). Sequential Changepoint Detection in Quality-Control and Dynamical Systems. *Journal of the Royal Statistical Society B,* **57** 613-658.

LAWSON, A. and RODEIRO, C. (2004). Developements in general and syndromic surveillance for small area health data. *Journal of Applied Statistics,* **31** 397-406.

LAWSON, A. B. and KLEINMAN, K. (Eds.). (2005). *Spatial and Syndromic Surveillance for Public Health*. New York: Wiley.

LOWRY, C. A., WOODALL, W. H., CHAMP, C. W. and RIGDON, S. E. (1992). A multivariate exponentially weighted moving average control chart. *Technometrics,* **34** 46-53.

MARSHALL, J. B., SPITZNER, D. J. and WOODALL, W. H. (2007). Use of the local Knox statistic for the prospective monitoring of disease occurrences in space and time. *Statistics in Medicine,* **26** 1579-1593.

PAUL, M., L. HELD, AND A. M. TOSCHKE. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine,* **27** 6250-6267.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, Chichester.

ROGERSON, P. A. (1997). Surveillance systems for monitoring the development of spatial patterns. *Statistics in Medicine,* **16** 2081-2093.

ROGERSON, P. A. (2001). Monitoring point patterns for the development of space-time clusters. *Journal of the Royal Statistical Society A,* **164** 87-96.

RYAN, T. P. (2000). *Statistical methods for quality improvement*. Wiley, New York.

SCHIÖLER, L. (2010). *Modelling the spatial patterns of influenza incidence in Sweden* (No. 2010:1). Gothenburg: Statistical Research Unit, Department of Economics, University of Gothenburg, Sweden.

SCHIÖLER, L. and FRISÉN, M. (2008). *On statistical surveillance of the performance of fund managers* (No. 2008:4): Statistical Research Unit, Department of Economics, University of Gothenburg, Sweden.

SHMUELI, G. and BURKOM, H. S. (2010). Statistical Challenges Facing Early Outbreak Detection in Biosurveillance. *Technometrics,* **52** 39-51.

SONESSON, C. (2007). A CUSUM framework for detection of space-time disease clusters using scan statistics. *Statistics in Medicine,* **26** 4770-4789.

SONESSON, C. and BOCK, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society A,* **166** 5-21.

SONESSON, C. and FRISÉN, M. (2005). Multivariate surveillance. In *Spatial surveillance for public health (*A. LAWSON and K. KLEINMAN, eds.) 169-186. Wiley, New York.

SULLIVAN, J. H. and JONES, L. A. (2002). A self-starting control chart for multivariate individual observations. *Technometrics,* **44** 24-33.

TARTAKOVSKY, A. G. and VEERAVALLI, V. V. (2008). Asymptotically Optimal Quickest Change Detection in Distributed Sensor Systems. *Sequential Analysis,* **27** 441 - 475.

WESSMAN, P. (1998). Some Principles for surveillance adopted for multivariate processes with a common change point. *Communications in Statistics - Theory and Methods,* **27** 1143-1161.

WOODALL, W. H. (2006). The Use of Control Charts in Health-Care Monitoring and Public-Health Surveillance. *Journal of Quality Technology,* **38** 89-134.

WOODALL, W. H. and MONTGOMERY, D. C. (1999). Research Issues and Ideas in Statistical Process Control. *Journal of Quality Technology,* **31** 376-386.

ZHOU, H. and LAWSON, A. B. (2008). EWMA smoothing and Bayesian spatial modeling for health surveillance. *Statistics in Medicine*, **27** 5907-5928.
ZHOU, Q., LUO, Y. and WANG, Z. (2010). A control chart based on likelihood ratio test for detecting patterned mean and variance shifts *Computational Statistics & Data Analysis*, **54** 1634-1645.

L. SCHIÖLER
M. FRISÉN
STATISTICAL RESEARCH UNIT
UNIVERSITY OF GOTHENBURG
SE 405 30 GOTHENBURG
SWEDEN
E-MAIL: linus.schioler@statistics.gu.se
marianne.frisen@statistics.gu.se

Research Report

2007:11	Frisén, M. & Andersson, E.	Semiparametric surveillance of outbreaks.
2007:12	Frisén, M., Andersson, E. & Schiöler, L.	Robust outbreak surveillance of epidemics in Sweden.
2007:13	Frisén, M., Andersson, E. & Pettersson, K.	Semiparametric estimation of outbreak regression.
2007:14	Pettersson, K.	Unimodal regression in the two-parameter exponential family with constant or known dispersion parameter.
2007:15	Pettersson, K.	On curve estimation under order restrictions.
2008:1	Frisén, M.	Introduction to financial surveillance.
2008:2	Jonsson, R.	When does Heckman's two-step procedure for censored data work and when does it not?
2008:3	Andersson, E.	Hotelling´s T2 Method in Multivariate On-Line Surveillance. On the Delay of an Alarm.
2008:4	Schiöler, L. & Frisén, M.	On statistical surveillance of the performance of fund managers.
2008:5	Schiöler, L.	Explorative analysis of spatial patterns of influenza incidences in Sweden 1999—2008.
2008:6	Schiöler, L.	Aspects of Surveillance of Outbreaks.
2008:7	Andersson, E & Frisén, M.	Statistiska varningssystem för hälsorisker
2009:1	Frisén, M., Andersson, E. & Schiöler, L.	Evaluation of Multivariate Surveillance
2009:2	Frisén, M., Andersson, E. & Schiöler, L.	Sufficient Reduction in Multivariate Surveillance
2010:1	Schiöler, L	Modelling the spatial patterns of influenza incidence in Sweden