On-line detection of onset of an outbreak – a semiparametric approach

E Andersson[1,2] and M Frisén[1]

[1] Statistical research unit, University of Gothenburg,
PO Box 640, SE 405 30 Goteborg, Sweden
(e-mail: marianne.frisen@statistics.gu.se)

[2] Department of occupational and environmental medicine, Sahlgrenska University Hospital and Sahlgrenska Academy, University of Gothenburg,
PO Box 414, SE 405 30 Goteborg, Sweden
(e-mail: eva.andersson@statistics.gu.se)

Abstract. Detection of a progressive increase as soon as possible after the onset of the outbreak is of interest. A semiparametric method is constructed to detect a change from a constant level to a monotonically increasing function. It is applied to Swedish incidence data. The suggested method is compared with subjective judgments as well as with other algorithms. The conclusion is that the method works well.

Keywords. Statistical surveillance, Semi-parametric, Outbreak, Likelihood ratio.

1 Introduction

To monitor an on-going process in order to detect a change is important in many areas: industrial process control, environmental monitoring and public health surveillance. In public health surveillance we can be interested in the time of the outbreak, for example of influenza. (Farrington, Andrews, Beal, et al. (1996)) have suggested a parametric surveillance method for influenza, where the in-control model is estimated from previous data. This works well for detecting an influenza season that is different from the average season. It will also indicate an outbreak that starts earlier than usual, but it will not call an alarm for the outbreak of a late season. We will suggest and evaluate a method which is suitable for outbreak detection also for seasons which do not agree with the average season.

Surveillance is detection of deviation from an in-control state. In order to detect the change of importance, we use a surveillance system consisting of an alarm statistic and an alarm limit. The monitored process is denoted X. At each decision time, s, we want to discriminate between two events, here denoted C (the change has occurred) and D (the change has not occurred yet). It has been shown that alarm systems based on likelihood ratios (between C and D) are optimal. The likelihood ratio (between C and D) is

$$\frac{f(X_s|C)}{f(X_s|D)},$$

where Xs={X(1), X(2), ..., X(s)}. The partial likelihood ratios are

$$\text{L(s,t)}=\frac{f(x_s|\tau=t)}{f(x_s|\tau>t)}.$$

The full likelihood ratio is the weighted sum of the partial likelihood ratios

$$w_1 \cdot L(s,1) + w_2 \cdot L(s,2) + ... + w_s \cdot L(s,s)$$

where wj= $P(\tau = j) / P(\tau \le s)$.

The full likelihood ratio is optimal in the sense that it minimized the expected delay for a fixed false alarm probability (Shiryaev (1963)). The maximal likelihood ratio is optimal in a minimax sense (Moustakides (1986)).

The change that we want to detect can be a step-change (shift in mean, variance or autocorrelation) or it could be a gradual change, e.g. a turning-point (change from expansion to peak to recession) or the onset of an outbreak (change from a constant level to an increasing function). Many surveillance methods are based on the parameters of the in-control and out-of-control states being known or possible to estimate. In many situations, especially with complex changes (e.g. a turning point or an outbreak) these parameters might be difficult to estimate with certainty. Therefore non-parametric solutions are of interest.

For outbreak detection Martìnez-Beneito, Conesa, López-Quílez, et al. (2008) suggest to monitor the differentiated series (Xt-Xt-1) in order to detect a change from white-noise process to an autoregressive process of order 1. However, to differentiate data means that information is lost and the dependency structure is changed. Mei (2008) discusses different ways to handle an unknown baseline at surveillance. We have developed a semi-parametric outbreak detection method, using the generalized likelihood ratio, GLR. This technique was used in previous work for turning point detection in business cycles: the cycles are very irregular over time, both in amplitude and length (Andersson (2006)) and thus it is difficult to find a model which is valid over time. Therefore a semi-parametric approach, based on monotonicity restrictions, was suggested and evaluated, see Andersson (2002) and Andersson, Bock and Frisén (2006). The vulnerability of a parametric model for the cycles was especially noticable for the situation when the in-control state (the pre-turn slope) was mis-specified, see Andersson, Bock and Frisén (2005).

In Section 2 we will describe the semiparametric method for outbreak detection, In Section 3 the method is evaluated and concluding remarks are given in Section 3.

2 Semiparametric method for outbreak detection

Our semiparametric method is parametric with respect to distribution as we use the regular exponential family but it is nonparametric with respect to the regression.

At the start of an outbreak the incidence is characterized by a change from a constant level to an increasing function. This was the start for the development of a semiparametric system for outbreak detection. Say that the outbreak starts at time τ . Then the expected incidence can be described as

$$\mu(0) = ... = \mu(\tau - 1) < \mu(\tau) \le ... \le \mu(s)$$

The monotonicity restriction contains two parts

$$\mu(0) = ... = \mu(\tau - 1)$$

and

$$\mu(\tau - 1) < \mu(\tau) \le ... \le \mu(s).$$

2.1 Maximum likelihood estimator of the order restricted regression

Frisén, Andersson and Pettersson (2009) give the maximum likelihood estimator of the outbreak regression described above for the exponential family. For a specific value τ the estimator is constructed by first computing a provisional series such that

$$Y^{\tau}(t) = \begin{cases} \sum_{j=0}^{\tau-1} X(j) / \sum_{t=0}^{\tau-1} 1, & t < \tau \\ X(t), & t \le \tau \end{cases}$$

The next step is to consider the second condition:

$$\hat{\mu}^{\tau}(t) = g(t \,|\, Y^{\tau}(0), Y^{\tau}(1), ..., Y^{\tau}(s))$$

where the function g(t) is the least squares estimator of the provisional series under the second monotonicity restriction. The estimator can also be seen as a pool-adjacent-violators algorithm (PAVA, see Robertson, Wright and Dykstra (1988)).

For certain distributions the least squares estimators given above are also maximum likelihood estimators, for example the normal distribution and the Poisson distribution.

2.2 A semi-parametric surveillance system

We now have a maximum likelihood estimate of the outbreak regression. The next step is to derive a surveillance method where these ML-estimates are used. In Frisén and Andersson (2009) the generalized likelihood ratio approach is used to derive an outbreak detection method, according to which an alarm is called when

$$\frac{\max f(x_s \,|\, C1)}{\max f(x_s \,|\, D)} > \text{k},$$

where k is the alarm limit. For a normal distribution, the maximum likelihood alarm statistic becomes

$$\frac{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{s} (x(i) - \hat{\mu}^{C1}(i))^2 \right)}{\exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{s} (x(i) - \hat{\mu}^{D}(i))^2 \right)},$$

where $\hat{\mu}^{C1}$ and $\hat{\mu}^{D}$ are the maximum likelihood estimates under monotonicity restrictions $\tau = 1$ and $\tau > s$. This alarm statistic is denoted OutbreakN.

In most public health applications, the Poisson distribution can be considered more interesting than the normal distribution. For Poisson, the maximum likelihood alarm statistic becomes

$$\prod_{t=1}^{s} \left(\frac{\hat{\mu}^{C1}(t)}{\hat{\mu}^{D}(t)} \right)^{x(t)},$$

which is denoted OutbreakP.

3 Evaluation of the outbreak detection system

In Frisén and Andersson (2009), the OutbreakP method was evaluated in a simulation study. Data were generated from a model that mimics Swedish influenza data. A Poisson distribution for these data was suggested in Andersson, Bock and Frisén (2008). Observations were generated according to the following model

$$X(t) \sim \begin{cases} Poi(\mu_0), & t < \tau \\ Poi(\mu(t)), & t \geq \tau \end{cases}$$

where Poi(*) refers to the Poisson distribution. The level μ_0 was roughly estimated to μ_0=1 from Swedish data for eight years. In Andersson, Bock and Frisén (2008) it was found that an exponential curve works well to mimic the increasing phase (for t$\geq\tau$), so that $\mu(t) = \exp(\beta_0 + \beta_1 \cdot (t - \tau + 1))$. The parameters were estimated to β_0=-0.26 and β_1=0.826.

3.1 Evaluation measures in surveillance

In surveillance we need evaluation measures that reflect the timeliness and the confidence of the alarms. When an alarm is called it can be a false alarm or a motivated alarm. The alarm limit can be set so that the false alarm property of the method is known. Examples of false alarm properties are the probability of a false alarm and the average run length to the first false alarm. A metric that mirrors the relation between the false alarms and the motivated ones is the predictive value (suggested by Frisén (1992)), i.e. the probability that the change has occurred given that an alarm is called:

$$\text{PV(t)} = P(C \mid t_A = t)$$

Note that PV is not necessarily constant over time.

In Frisén and Andersson (2009) the alarm limit in the simulation study of OutbreakP is set so that the predictive value is higher than 0.99 at all time points, at least up to t=20.

An important aspect for a surveillance system is a quick detection, i.e. a short time between the change and the alarm. This can be measured by the conditional expected delay

$$\text{CED(t)} = E\left[t_A - \tau \mid t_A \geq \tau, \tau = t \right]$$

For many methods CED is not constant over t and it can be important to study the whole delay curve. In Frisén the delay of the OutbreakP method differed depending on τ, so that CED(1)=3.1 and CED(10)=2.2 and CED(20)=2.0. Thus, the delay is longest for early onsets. This is reasonable since we have no information regarding the baseline or the outbreak, but all information comes from data, which is sparse early on in the surveillance.

3.2 Comparison to subjective judgement

In Frisén, Andersson and Schiöler (2009) the semiparametric OutbreakP method is compared to the subjective judgments by twenty-six medically trained individuals. The subjective method was less efficient than the OutbreakP method. However, the main disadvantage for the subjective method turned out to be the large variation between the individuals.

3.3 Comparison to the Shewhart method

In Frisén and Andersson (2009) a robustness study was conducted, in which the parametric Shewhart method was compared to the semi-parametric OutbreakN method. Thus in this comparison the assumption of a normal distribution was used. Observations were generated according to the following model

$$X(t) \sim \begin{cases} N(\mu_0; \sigma), & t < \tau \\ N(\mu(t); \sigma), & t \geq \tau \end{cases}$$

where μ_0=20, σ=10 and $\mu(t) = \exp(\beta_0 + \beta_1 \cdot (t - \tau + 1))$ and β_0=2.67, β_1=0.68. The robustness of the Shewhart method to a mis-specification was evaluated, i.e. i.e. the Shewhart method in which the wrong baseline (μ_0 + a) was used. A 95% confidence interval for μ_0 gave the limits 16 and 24 and these values were used as the mis-specified values of the in-control level.

Table 1: Delay when the outbreak occurs at time τ , CED(τ)

	Shewhart correct	Shewhart mis-spec (too high)	Shewhart mis-spec (too low)	OutbreakN
τ =1	0.9	1.0	0.7	1.6
τ =5	0.9	1.0	0.7	1.0
τ =10	0.9	1.0	0.7	0.9

Table 2: Predictive value of an alarm at time t, PV(t)

	Shewhart correct	Shewhart mis-spec (too high)	Shewhart mis-spec (too low)	OutbreakN
t=1	0.3	0.4	0.3	0.8
t=5	0.8	0.9	0.6	0.8
t=10	0.8	0.9	0.6	0.8

The conclusions from the tables are that when the baseline is overestimated, the CED is hardly better than for the nonparametric method and when the baseline is underestimated the PV is very low. Thus, uncertainty about the baseline will mean that the properties of the method are highly uncertain.

4 Conclusion

We have suggested a definition of outbreak which can be useful. This definition does only involve the monotonic increase and not the level of the baseline. Our method is thus nonparametric with respect to the regression. An advantage with knowledge of a parametric model is that more information is available and consequently a drawback of any non-parametric estimation is that the only information comes from data. Thus for very few data points (early on in the surveillance) the ability to detect changes is bound to be rather poor. However, for later changes the semi-parametric approach will work well since then data provide enough information and we avoid the traps of the mis-specified parametric approach. Our conclusion by the simulations and experimental evaluations and by application of the method to Swedish data of some diseases is that the method is promising.

References

Andersson, E. (2002), Monitoring Cyclical Processes - a Nonparametric Approach, *Journal of Applied Statistics*, **29**, 973-990.

Andersson, E. (2006), Robust on-Line Turning Point Detection. The Influence of Turning Point Characteristics In: *Frontiers of Statistical Quality Control*, Vol. 8 (H.-J. Lenz and P.-T. Wilrich, Eds) Warsaw: Physica Verlag, 223-248.

Andersson, E., Bock, D., and Frisén, M. (2005), Statistical Surveillance of Cyclical Processes. Detection of Turning Points in Business Cycles, *Journal of Forecasting*, **24**, 465-490.

Andersson, E., Bock, D., and Frisén, M. (2006), Some Statistical Aspects on Methods for Detection of Turning Points in Business Cycles, *Journal of Applied Statistics*, **33**, 257-278.

Andersson, E., Bock, D., and Frisén, M. (2008), Modeling Influenza Incidence for the Purpose of on-Line Monitoring, *Statistical Methods in Medical Research*, **17**, 421-438.

Farrington, C. P., Andrews, N. J., Beal, A. D., and Catchpole, M. A. (1996), A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease, *Journal of the Royal Statistical Society A*, **159**, 547-563.

Frisén, M. (1992), Evaluations of Methods for Statistical Surveillance, *Statistics in Medicine*, **11**, 1489-1502.

Frisén, M., and Andersson, E. (2009), Semiparametric Surveillance of Outbreaks, *Sequential Analysis*, in press.

Frisén, M., Andersson, E., and Pettersson, K. (2009), Semiparametric Estimation of Outbreak Regression, *Statistics*, in press.

Frisén, M., Andersson, E., and Schiöler, L. (2009), Robust Outbreak Surveillance of Epidemics in Sweden, *Statistics in Medicine*, **28**, 476-493.

Martìnez-Beneito, M., Conesa, D., López-Quílez, A., and López-Maside, A. (2008), Bayesian Markov Switching Models for the Early Detection of Influenza Epidemics, *Statistics in Medicine*, **27**, 4455-4468.

Mei, Y. (2008), Is Average Run Length to False Alarm Always an Informative Criterion?, *Sequential Analysis*, **27**, 354-376.

Moustakides, G. V. (1986), Optimal Stopping Times for Detecting Changes in Distributions, *The Annals of Statistics*, **14**, 1379-1387.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), Order Restricted Statistical Inference, Chichester: Wiley.

Shiryaev, A. N. (1963), On Optimum Methods in Quickest Detection Problems, *Theory of Probability and its Applications*, **8**, 22-46.