ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Prospective Statistical Surveillance in Public Health

By

Nikolas {John} Karatzas

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece
May 2012


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

Προοπτική Στατιστική Παρακολούθηση στη Δημόσια Υγεία

Νικόλαος Ιωάν. Καρατζάς

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα

Μάιος 2012


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

DEDICATION

This study is dedicated to my parents and my beloved brother.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ACKNOWLEDGEMENTS

This study is based upon the paper of Sonesson and Bock(2003) with title “A review and discussion of prospective statistical surveillance in public health”. General guidance and helpful comments were received by my supervisor Stelios Psarakis(Associate Professor at the Department of Statistics of the Athens University of Economics and Business).


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

VITA

I was born in Athens in 1985. I started my BSc in Statistics at the year 2003 when I entered Department of Statistics of the Athens University of Economics and Business. At the October of the year 2010, I started my MSc in Applied Statistics in the same University.

During my BSc studies, I had the opportunity to carry out my practice in Statistics in SPSS B.I. Greece and see how this field is applied to a professional environment. There, I gained important knowledge about algorithms, programming languages (e.g. Python) and other statistical packages (e.g. SPSS Modeler). I also worked in ISS as a Customers' Service Operator, representing the telecommunication's company "On Telecoms".

After having my graduation from BSc in Statistics, I worked in SPSS B.I. Greece as a data analyst at the consulting department. My duties were analysing data, organising databases and developing statistical algorithms for that cause.

Finally, at the year of 2011 I worked in Hellenic Statistical Authority as an Area Manager for the purpose of the 2011 population census of Greece.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ABSTRACT

Nikolas Karatzas

"Prospective Statistical Surveillance in Public Health"

May 2012

Current research is a study of public health surveillance and refers to the applications of statistical quality control for that purpose. Sonesson and Bock(2003) presented a review paper and we are based upon this study. We develop the methods presented in this paper. Simultaneously, we present simulations and the theoretical background behind these methods. Finally, as Sonesson and Bock presented bibliography until the year 2003, we are interested in presenting bibliography for the period 2003-2012.

At this study we give an idea of how we may apply the main principles of Statistical Quality Control to the surveillance of public health. Methods such as the CUSUM scheme and measures such as the Average Run Length and the expected delay are used for the construction of surveillance systems in the public health field. The optimality of these systems is an important factor since the public health surveillance demands the best results from us.

These systems' purpose is to detect peaks in the mean number of "events" in which case we should have an alarm. As a consequence, the purpose of these systems is to detect possible epidemics and through a prospective view we should be warned in order to proceed in the appropriate preventive actions.

We make a separation to our methods based on the assumption according to if the incidences are following a Poisson process or not. We also have three types of systems which are based on the temporal, spatial and spatial-temporal surveillance factors. For each case of these factors, we analyse different methods.

ΠΕΡΙΛΗΨΗ

Νικόλαος Καρατζάς

«Προοπτική Στατιστική Παρακολούθηση της Δημόσιας Υγείας»

Μάιος 2012

Η παρούσα εργασία είναι μια μελέτη με αντικείμενο την παρακολούθηση της δημόσιας υγείας και αναφέρεται στις εφαρμογές του στατιστικού έλεγχου ποιότητας για αυτό το σκοπό. Οι Sonesson & Bock(2003) παρουσίασαν μια κριτική εργασία και βασιστήκαμε σε αυτή τη μελέτη. Αναπτύσσουμε τις μεθόδους που παρουσιάζονται σε αυτή την εργασία. Ταυτόχρονα παρουσιάζουμε προσομοιώσεις και το θεωρητικό υπόβαθρο αυτών των μεθόδων. Τέλος, ενδιαφερόμαστε στην παρουσίαση βιβλιογραφίας για την περίοδο 2003-2012 από τη στιγμή που η μελέτη των Sonesson & Bock δίνει βιβλιογραφία μέχρι το έτος 2003.

Σε αυτή τη μελέτη δίνουμε μια ιδέα για το πώς μπορούμε να εφαρμόσουμε τις βασικές αρχές του Στατιστικού Έλεγχου Ποιότητας στην παρακολούθηση της δημόσιας υγείας. Μέθοδοι όπως το διάγραμμα CUSUM και μέτρα όπως το ARL και η αναμενόμενη καθυστέρηση μέχρι να έχουμε συναγερμό(ED) χρησιμοποιούνται για την κατασκευή συστημάτων παρακολούθησης στο πεδίο της δημόσιας υγείας. Η μέγιστη λειτουργία αυτών των συστημάτων είναι ένας σημαντικός παράγοντας αν λάβουμε υπόψη ότι η παρακολούθηση της δημόσιας υγείας απαιτεί από εμάς, τα καλύτερα δυνατά αποτελέσματα.

Ο σκοπός αυτών των συστημάτων είναι να ανιχνεύουν αλλαγές στο μέσο αριθμό κρουσμάτων, περίπτωση στην οποία θα έπρεπε να έχουμε συναγερμό. Κατά συνέπεια, ο σκοπός αυτών των συστημάτων είναι να ανιχνεύουν πιθανές επιδημίες και με μια προοπτική οπτική θα έπρεπε να προειδοποιηθούμε ούτως ώστε να πάρουμε τα κατάλληλα μέτρα αποτροπής μιας ανάλογης κατάστασης.

Κάνουμε ένα διαχωρισμό στις μεθόδους μας με βάση το αν ακολουθούν ή όχι την Poisson διαδικασία. Επίσης, έχουμε 3 τύπους συστημάτων για την παρακολούθηση της δημόσιας υγείας βασιζόμενοι στους παράγοντες του χρόνου, του χώρου και του χωρο-χρόνου. Για κάθε έναν από αυτούς τους παράγοντες, αναλύουμε διαφορετικές μεθόδους.

TABLE OF CONTENTS

LIST OF TABLES


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

LIST OF FIGURES

Chapter 1
Introduction

Statistical process control has its foundations in the industrial field and has been used with different ways and methods for the extraction of useful conclusions and prevention of unpleasant situations (e.g. the production of dysfunctional products). For that purpose statistical quality control has been widely spread the last thirty years and that is the reason why it has been adopted and developed especially by the states whose economy is based in the industrial production.

On the other hand, and beyond the financial motive, this field has been widely spread in the academic community. The charts of Shewhart have been just the beginning of the development of some new methods such as the Cumulative sum methods and the Exponential weighted moving average method. A huge amount of questions have been answered and even greater is the number of new ideas which have been generated.

The idea of the statistical process control is used nowadays in different fields and for different reasons. The construction of systems which give us the opportunity to proceed in the surveillance of some factors is the main purpose of this development. For example, in the financial sector someone may be interested in the stock market and the surveillance of the price of some stocks. Actually, we would like to have a system, where the prices of the stocks might be monitored in order to take preventive measures (such as an investment) if their course is out-of-control.

In this study, we are interested in the statistical process control applied in the field of public health. Public health is a very sensitive issue not only for the individuals but also for the communities. Monitoring the public health is a big challenge as there are numerous factors (e.g. delay of an alarm, delay of reporting, not accurate records etc.) that can provoke for example an outbreak of an epidemic. Furthermore, constructing a system for a health natured problem, means taking a huge risk as what is at stake (e.g. at the case of an epidemic) are

peoples' lives. We have to be precise in our calculations, quick in our detections and confident for our conclusions.

Sonesson and Bock(2003)[140] presented a definition for public health surveillance that we believe that it is representative: *"Public health surveillance is the on-going systematic collection, analysis and interpretation of outcome-specific data that are essential to the planning, implementation and evaluation of public health programs, closely integrated with the timely dissemination of these data to those who are responsible for prevention and control"*.

Imagine now all the available data that exist in the different organizations, hospital records, private doctors' records etc. The amount of data which we can analyze is large. We can tackle with the problem of the data collected, in two different ways in order to exploit all the information we can. The first is the retrospective surveillance. In this case the data are fixed and we proceed in the analysis of all the available data at the same time. The second way is the prospective surveillance or on-line monitoring. In this case we are interested to analyze the data through time and at different time points. We proceed in a repeated analysis of data accumulating over time and we make some decisions at some time points about the state of our process. The timeliness of these decisions (for simplicity we use discrete time) is essential for our analysis since we want to detect an increase in the number (or the rate) of incidences (e.g. a disease or car accidents etc.) as quickly as possible in order to proceed in the appropriate actions. In this paper we discuss on the case of the prospective surveillance since it is a more complicated and more interesting situation from a statistical point of view.

For the on-line public health surveillance we have to consider some factors that do not exist in the industrial process control. These are for example the time delay caused by the delay of the reporting an incident, the seasonal effect of a disease in the population, lack of accuracy, missing reporting data, the biased reporting, the false or the delay of the diagnosis etc. As we can easily understand, public health surveillance may include such factors which make our role more difficult but as we discuss in this study, there are several ways to overcome such disadvantages of our process. Epidemiological data need special treatment and their features are special. That is the reason why we have to deal with them in a special way.

In this study we focus, though, more in the statistical part of the prospective surveillance of public health. That is not the epidemiological problems mentioned above but the statistical solutions and methods which lead us to conclusions about the state of our process. Thus, the main purpose of this study is to examine different methodologies for the prospective case of public health surveillance and evaluate them through different measures.

1.1 Issues of the Public Health Surveillance

Such a measure might be the Average Run Length which is used mostly in the industrial process control. For the monitoring of public health surveillance we are dealing with other measures such as the temporal delay, the probability of detecting a true alarm, the detection of a change in the mean in a specified limited time delay etc. These measures are presented in chapter 3 of this study and can be used to measure the reliability of our results or for the comparison between methods.

In order to construct a surveillance method we need to specify some factors. These are the *alarm statistic* and the *alarm limits*. Each time we are interested in developing a surveillance method, we have to choose these factors and deal with the proper properties of the system often expressed in terms of an optimality criterion. This is another issue which is given in chapter 4.

In order to make our system an optimal one we may use these measures. For example measures such as the delay, which we mentioned above, can be used in order to construct an optimal criterion which will lead us to an optimal system. It is rational for someone to think that the shorter the delay the better is the method. A temporal delay between the alarm and the real out-of-control state is almost certain to happen but the minimization of such a delay will give us the opportunity to take the appropriate actions at the shortest time interval. The time delay between the system's alarm and the real change of a process (e.g. an increase in the rate of a disease) can be of interest and its minimization is our main goal. That is the case for the most of the methods in public health since the time delay is of special interest. This measure is significant in this

case if we consider the fact that a long delay might mean an outbreak of an epidemic and maybe the loss of lives.

1.2 Structure of the Study

In the second chapter we mention some general concepts of the SPC applied in the public health surveillance and we give a full description of them as long as a symbolism necessary for the better understanding of this study.

In the third chapter we give some measures of evaluation which we use in the public health in order to compare different monitoring systems and which gives us an overall view of the performance of our surveillance.

In the fourth chapter we mention some issues as far as we are concerned about the optimality of a method. Three optimal criteria and a general description and criticism for their function are given. In this study the criterion of the minimum expected delay is of special interest.

Our study consists of three main parts which are referred to different cases of our surveillance. The first two parts are referred to the cases of the Poisson or non Poison processes (processes with time dependencies) and in these parts we describe different methods according to their relation with the factor of time. We could say that in these first two parts we are dealing with the temporal surveillance and in the third part we present the case of spatial surveillance. With such a way we manage to give a complete view of the aspects of the on-line surveillance in public health. Space and time are the two “dimensions” from which our methods are constructed.

In the fifth chapter we give two methods for the case of using the time between events to study the Poisson process. These are the Sets method and the Cuscore method.

In the sixth chapter we present the case of using the number of events to study the Poisson process. On the contrary from the previous chapter, in this case the time intervals are fixed and are not used to be corresponded with the incidences. The Poisson CUSUM method is described thoroughly and the case of the alarm statistic to be the maximum value of the conditional likelihood ratios is given descriptively due to the complicated calculations which did not

allow us to run a simulation. However, the study of Lie (1993)[84] is more than enough if the reader wants more information about this issue.

The next method which is described in the seventh chapter, is developed through weekly reports from the CDC. That is the Historical Limits method. This is a method which is constructed in a way to tackle with the seasonality of some diseases using historical data to set the limit of the process.

Until this point of the study, we had to do with suboptimal Poisson processes. In the last chapter of the first part (chapter 8) we present the case of likelihood ratio and the Shiryaev-Roberts ratio methods which are optimal Poisson processes.

In public health surveillance it is common to assume a Poisson process for the cases of disease but when this assumption is not appropriate, complicated time dependent processes have been used. In the second part which consists of the 9^{th} chapter, we present different issues and models which are dealing with processes with time dependencies. A general description and bibliography is given for each case if the reader is interested to examine these issues more thoroughly. This is a huge scientific section and the evolution of different models in order to develop reliable systems of monitoring, seems to be uncountable. On the other hand, the calculations make the construction of such models very difficult and complex.

In the last part of our study which is given in chapter 10, we present the case of spatial surveillance. This field seems to gain the epidemiologist's interest the last few years, since the interest is moved to the clustering of diseases or geographical patterns etc. Considering the amount of epidemiological books which are referred in the spatial feature of surveillance, we realize that this is going to be a highly developing sector in the future.

Finally, some thoughts are given in the last chapter in order to make a general conclusion of our study and give some directions for further investigations in the future.

1.3 General Issues

To make this study a reality we are influenced by the paper of Sonesson and Bock who gave us a review of the prospective surveillance in public health until the year 2003. Several cases and methods of public health are summarized in a few pages. This is a paper which is strongly recommended for the reader if they want to have a good knowledge about the issues raised in this field. We are based on this paper for our study and we present further literature on the issues mentioned in Sonesson and Bock(2003)[140] for the period 2003-2012.

We tried to make as much calculations as possible in order to present reliable and visual conclusions to the reader through simulations. We tried to give an overall view of the public health issues and for this purpose we worked in the statistical packages of S-plus and Minitab. The code which is developed in these packages in order to have the results presented in the main part of our study is given to the Appendix. For the convenience of the reader we partitioned the Appendix in parts so that any confusion can be avoided. Furthermore, we describe in the code the steps we followed, so that its comprehension will be easy even for a reader who is not fully familiar with statistical packages such as S-plus.

Chapter 2
General Concepts

With the term "statistical surveillance" we mean the on-line monitoring of a stochastic process $X = \{X_t, t = 1,2,...\}$ with the aim of detecting an important change at an unknown time point $\boldsymbol{\tau}$, as quickly as possible. At each decision time point $\boldsymbol{s}$, we want to make a decision about the state of our process. We denote the in-control process by ***D(s)*** and the out-of-control process by ***C(s)***. The two states we are interested, can be expressed as follows:

$$D(s) = \{\tau > s\}$$
$$C(s) = \{\tau \leq s\}$$

In order to achieve this decision we have to use the accumulated observations $X_s = \{X_t, t \leq s\}$.If our process is in out-of-control state ***C(s)*** then we say that $X_s \in A(s)$ where ***A(s)*** is our alarm set in which case our system triggers an alarm.

For the statistical surveillance in public health we are interested in the time of an alarm. Our system triggers an alarm at a time point t_A where this time point can be expressed as follows:

$$t_A = \min\{p(X_s) > g(s)\},$$

where $p(X_s)$ is our alarm function and $g(s)$ is a control limit.

In most cases, the change is referred in a shift of the initial mean of the process. The random process that determines the state of the system is denoted by $\boldsymbol{\mu(t)}$ for $t=1,2,\ldots$ In other words, we assume our process has an initial mean value μ_0 for $t = 1,2,\ldots,\tau-1$ and at time point $t = \tau$ we have a change in the process and the level of the mean value μ_0 is moved and remains to a new level μ_1 for $t = \tau, \tau + 1,\ldots$.In this sense we can express the two states mentioned above, as follows:

$$D(s) = \{\mu(s) = \mu_0\}$$
$$C(s) = \{\mu(s) = \mu_1\}$$

The initial and the new level of μ are regarded as known values and the time τ where the change occurs is regarded as a random variable with the density:

$$\pi_t = P(\tau = t) \text{ and } \sum \pi_t = 1 - \pi_\infty$$

The intensity of a change is denoted by:

$$v_t = P(\tau = t \mid \tau \geq t)$$

When we have to specify a distribution for τ, we use the geometric distribution. This specification of the distribution of τ, is suitable when the intensity of a shift is constant for each time point.

For our simulations we make the assumption that our surveillance is stopped when we have an alarm. Only one alarm is possible in our cases. That is called *active* surveillance as it is defined in Frisén and de Maré(1991)[46]. In most of our cases in this study we are dealing with active surveillance problems. Additionally, we have to make the assumption that ***X(i)-μ(i)*** are independent random variables.

Table 2.1: Symbols for the general concepts of Public Health surveillance

General Concepts	***Symbolism***
Stochastic Process	$X = \{X_t, t = 1,2,...\}$
Time of Change	τ
Decision Time	s
In-Control Process	$D(s)$
Out-of-Control Process	$C(s)$
Alarm Set	$A(s)$
Time of an Alarm	t_A
Intensity of a Change	v_t

Chapter 3
Measures of Evaluation

In order to test our method's performance it is common to use the measures of the significance level and the power. The problem for these measures is that we do not take into account some important factors such as the dependence on the length of the period of surveillance and the exact time point that the change occurs. In order to consider such factors in our surveillance method we have to generalize the two measures mentioned above.

Before continuing to the presentation of these measures, it is vital, for us, to evaluate the statistical properties. For example there are procedures in which long series are used and others that we care for the performance not too long from the start. Generally, in order to evaluate a surveillance system we have to consider some important factors such as the availability of information. This could mean the time delay generated from the reporting process or the time delay generated from the process of identifying and confirming an outbreak of a disease.

Another problem is the simplicity of a method. This factor is usually determined objectively, according to the application of our surveillance. There are applications that the simplicity of a method is vital and we can just use the *Average Run Length (ARL)* tool. However, there are applications that a great amount of information is needed. In these applications we have to use more complex methods and more than one measure for our system's evaluation. An example here can be the supervision of the foetal heart rate during labour which was presented by Frisén(1992)[36]. If an abnormality has occurred it is important for us to detect the problem as soon as possible so that immediate actions will have to be taken such as a caesarean section. It is easily understandable that in this application we have to deal with the dependence of the time factor. Acting too late may have as a result the loss of a life.

Several studies are referred in such measures. An example might be the study of Andersson(2003)[3]. These measures of evaluation are commonly used in order to compare different methods. A graphical evaluation of such measures

are given in the paper of Frisén and Gottlow(2003)[45] for a constructed statistical program. There are other studies though which in order to compare different methods, proceed in the development of some other criteria. For example in Sego et al.(2008)[126] and Sego (2006)[125] it was used the steady-state average run length instead of the ARL.

In this chapter we are examining different types of measures for the purpose of evaluating methods in surveillance of public health.

3.1 Average Run Length (ARL)

A common measure of evaluation is the average run length until we have an alarm. If our process is in-control and we have an alarm then this alarm is false. The distribution of the false alarms is expressed by the following equation:

$$ARLo = E[t_A \mid \tau = \infty]$$

Approaching this issue from a general perspective, we have a rule which we use to extract the limits of our method and from which we define the in-control or out-of-control state of our process. That is:

$$ARLo = ARL$$

The *ARL* measure is used for the design of the control charts in different applications. Its disadvantages, though, are of vital significance. The *ARL* gives us limited information and is calculated considering the assumption that the surveillance begins at the same time when the change occurs. This assumption simplifies the problems for us in theory but in practice, the change occurs at an unknown time $\boldsymbol{\tau}$ since the surveillance started.

On the other hand this measure is common to be used in comparing different methods. Comparisons of methods with the same *ARL* is common in statistical surveillance as we can find out easily which is the quickest method in realising the change. These comparisons of the performance of different methods with the same *ARL* may depend on different ways on the value of τ. For

the average run length we assume $\tau=1$. The dependence of the average run length after the change at τ is demonstrated by another measure which we are presenting below (*Probability of Successful Detection*).

In the simulation presented below we have such a comparison between the Shewhart and the CUSUM method for the same average run length.

Simulation

In this simulation we are going to compare the methods of the CUSUM and the Shewhart test for the same *ARL*. Suppose we have a mean μ_0 for the time $t = 1, 2,\ldots, \tau-1$ and a mean μ_1 for the time $t = \tau,\tau+1,\ldots$.The time $\boldsymbol{t = \tau}$ is when the change in the mean occurs. We generate from Normal distribution 140 replicates with a mean of $\mu_0 = 0$ and 140 observations with a mean of $\mu_1 = 1{,}5$.The standard deviation is $\sigma = 1$ .So our change in the mean is $1{,}5\sigma$.We divide the observations for each case in 20 samples of size $n=7$ and we assume that we take each of these samples in the time periods $t=1,2,\ldots,40$. Practically that means that we have a time point of change at $\tau = 21$. We remind the statistic of the tabular CUSUM scheme for the case of an increase to be:

$$S_i^+ = \max\left(0, Y_i - (\mu_0 + k) + S_{i-1}^+\right)$$

Where k is given by the following:

$$k = \frac{|\mu_1 - \mu_0|}{2} = \frac{3}{4} = 0.75 \text{ for our case.}$$

The observations are given in the table below:

Table 3.1: Generated observations from N(0,1) and N(1.5,1)

Time	$\mu_0 = 0$	$\mu_1 = 1{,}5$
1	0,185573	1,75188
2	-0,247210	1,19523
3	0,355101	1,45971
4	0,357489	1,51561
5	0,333314	1,49897
6	-0,010791	1,24824
7	-0,580087	1,67122
8	0,205866	1,17702
9	0,081491	1,67445
10	-0,040827	1,53925
11	0,201591	1,46459
12	-0,914889	0,98769
13	-0,785749	1,07834
14	-0,134363	1,61910
15	0,326480	2,11732
16	-0,145937	1,41975
17	-0,343590	1,04650
18	-0,475557	1,83593
19	0,051910	1,04705
20	0,234598	1,28201

For a parameter limit $h=5\sigma$ for the CUSUM method and the shift in the mean mentioned above, we have the following CUSUM chart:

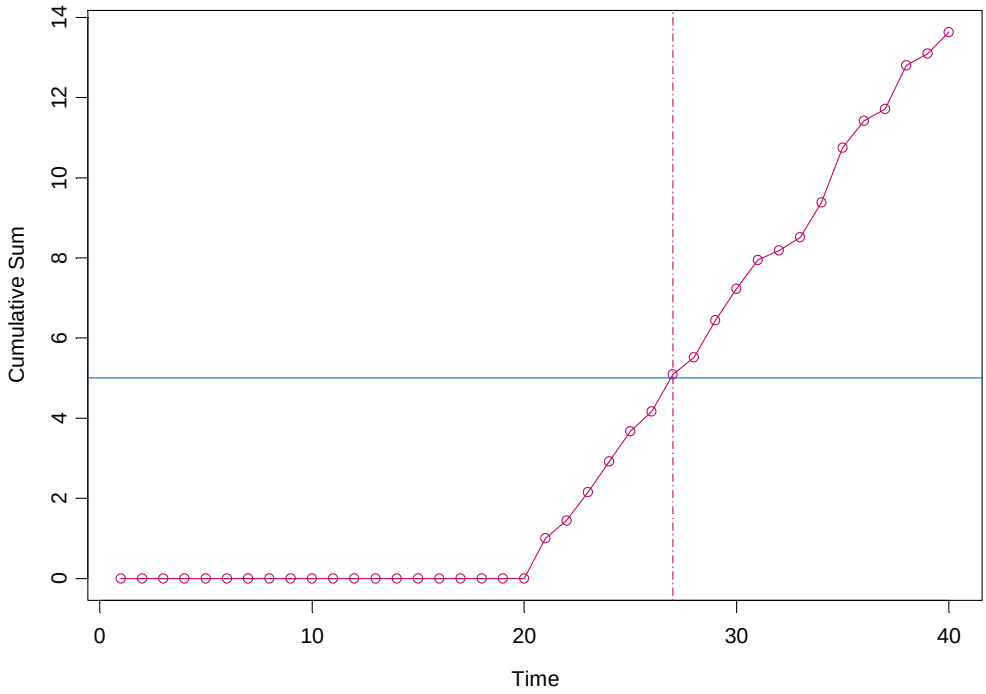


Figure 3.1: The CUSUM scheme for the 40 observations generated from Normal distribution and for a limit $h=5\sigma$

From the CUSUM's chart we see that the average run length is about 27 observations. Now for the same value of *ARL* we can figure out easily the limits which we are going to need for the Shewhart's chart:

$$ARL = \frac{1}{p},$$

where p is the probability one point plots out of control. Actually, we can symbolize this probability as follows:

$$p = P(|X_t| > h) \textit{ (for the two-sided method)}$$

and

$$p = P(X_t > h) \textit{ (for the one-sided method)}$$

With h we denote the limits of our chart. So, for our example we have:

$$p = \frac{1}{ARL} = \frac{1}{27} = 0.037 \Leftrightarrow$$
$$\Leftrightarrow P(X_t > h) = 0.037 \Leftrightarrow$$
$$\Leftrightarrow 1 - P(X_t \leq h) = 0.037 \Leftrightarrow$$
$$\Leftrightarrow 1 - \Phi(h) = 0.037 \Leftrightarrow$$
$$\Leftrightarrow \Phi(h) = 0.9630$$

So our limit's h value is about *1.79*. And so we have the following Shewhart chart for the observation of the mean:

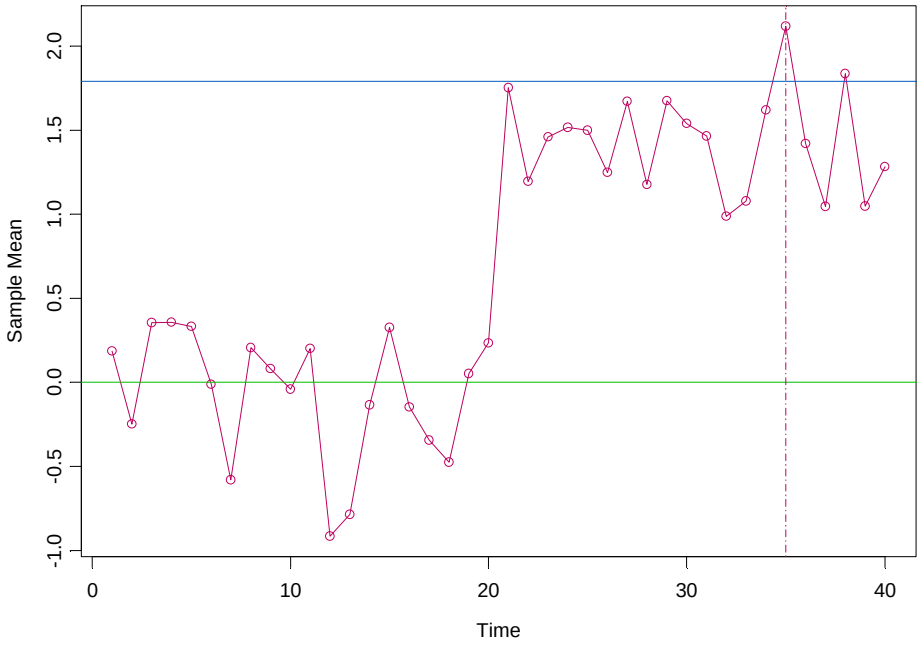


Figure 3.2: The one sided Shewhart scheme for the 40 observations generated from Normal distribution and a limit extracted from the CUSUM h=1.79

We conclude that the CUSUM's method triggers an alarm at the time of alarm $t_A = 27$ (or at the 27^{th} sample). On the other hand, based on the same *ARL* from the CUSUM's method, the system of the Shewhart's method triggers an alarm at the time $t_A = 35$.From these comparison we take that the CUSUM system is much better than the Shewhart method since it realizes the change in the mean much sooner with the same average run length. In the sections below we are presenting measures of evaluation based on the *time delay*. In terms of time delay the CUSUM's delay is $t_A - \tau = 6$.On the other hand the time delay

for the Shewhart scheme is larger: $t_A - \tau = 14$.This big difference for our case is natural since the CUSUM method is designed in a way that considers the cumulative sums of our observations. Thus, immediate and large changes in the mean are detected much sooner than the Shewhart's method.

3.2 Probability of a False Alarm (PFA)

In this measure we are dealing with the probability of a false alarm not later than a time *t* from the start. The probability of a false alarm actually can be interpreted as the *type I error* of a testing hypothesis. That is to reject the null hypothesis when we accept it. In our case an alarm is triggered when our process is not out-of-control. The interpretation with the previous measure of the Average Run Length for the null hypothesis (ARL_0) is the same.

Definition:
The *Probability of a False Alarm is the probability having an alarm when there is no real change in the process and it is given by the following expression:*

$$PFA(t) = P(t_A = \mathrm{t} \mid \tau > \mathrm{t})$$

This measure is a function of *t* and is denoted by a_t.For example for the Shewhart test we have: $a_t = 1 - (2\Phi(h) - 1)^t$ where Φ is the normal probability distribution function. From the example above we have $\Phi(h) = 0.9630$. Thus, for our example, the probability of a false alarm is shown in the figure 3.3:

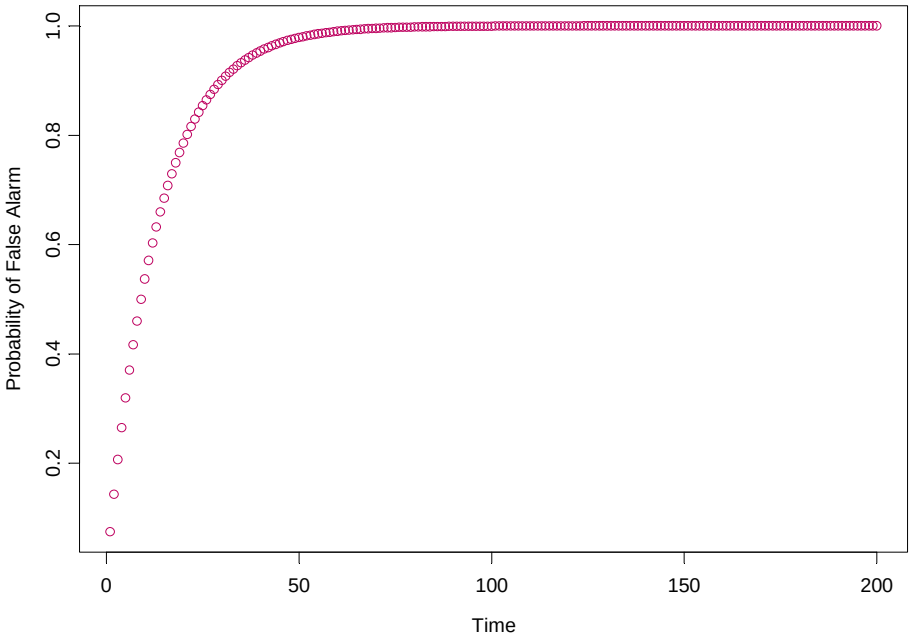


Figure 3.3: The visual projection of the PFA for the Shewhart scheme of the example in section 3.1

A summarizing measure of the false alarm distribution is the *total probability of a false alarm*:

$$P(t_A < \tau) = \sum_{t=1}^{\infty} P(\tau = t) P(t_A < t \mid \tau = t)$$

An assumption is needed at this point for the distribution of $\boldsymbol{\tau}$.Usually we assume that $\boldsymbol{\tau}$ ~ ***Geometric (p)***, where p is our parameter and reflects the probability we have an alarm. That is our "success". This assumption is appropriate when the intensity of a shift is constant for each time point. As a result, the first factor in the sum of the total probability of a false alarm does not depend on the method but only on the true intensity $\boldsymbol{v}$. The second factor depends only on the run length distribution when no change has occurred.

3.3 Probability of Successful Detection

A measure to demonstrate the dependence of the ARL after the change, on the time τ, is the *Probability of Successful Detection*. For this measure we are interested in the delay between the time when the change occurred and the time of our system's alarm ($t_A - \tau$).

An example would be the case of an infectious disease, where we have to detect its outbreak in a given time interval. If we do not detect the outbreak in time, we will not be able to prevent an epidemic. So only a limited time delay is permitted, since many lives may be at risk. If we take the appropriate preventive measures in a short time after a change, an outbreak will be avoided and that is the reason why this measure is so important. The delay is denoted by ***d.*** There are some applications that only a limited time delay can be tolerated, in which we can consider the *Probability of Successful Detection.*

Definition:
The *Probability of Successful Detection is the probability that the change is detected with a delay that is no longer than d, given that there was no alarm before the change. It is given by the following expression:*

$$PSD(d,t) = P(t_A - \tau \le d \mid t_A \ge \tau = t)$$

The *Probability of Successful Detection* is given for the Shewhart method and for a change in the mean from:

$$PSD(d,t,\mu(t)) = 1 - (\Phi(\mathrm{h} - \mu_1))^d .$$

In the previous example we had for the Shewhart method a time delay of about *d=14*. Thus, the probability of successful detection is:

$$PSD(d,t,\mu(t)) = 1 - (\Phi(\mathrm{h} - \mu_1))^d \Leftrightarrow$$
$$\Leftrightarrow PSD(d = 14, t = \tau, \mu(\tau) = 1.5) = 1 - [\Phi(1.79 - 1.5)]^{14} \Leftrightarrow$$
$$\Leftrightarrow PSD(d = 14, t = \tau, \mu(\tau) = 1.5) = 0.9989$$

The following graph gives us a visual image of the Probability of Successful Detection according to the time delay for our example's Shewhart scheme:

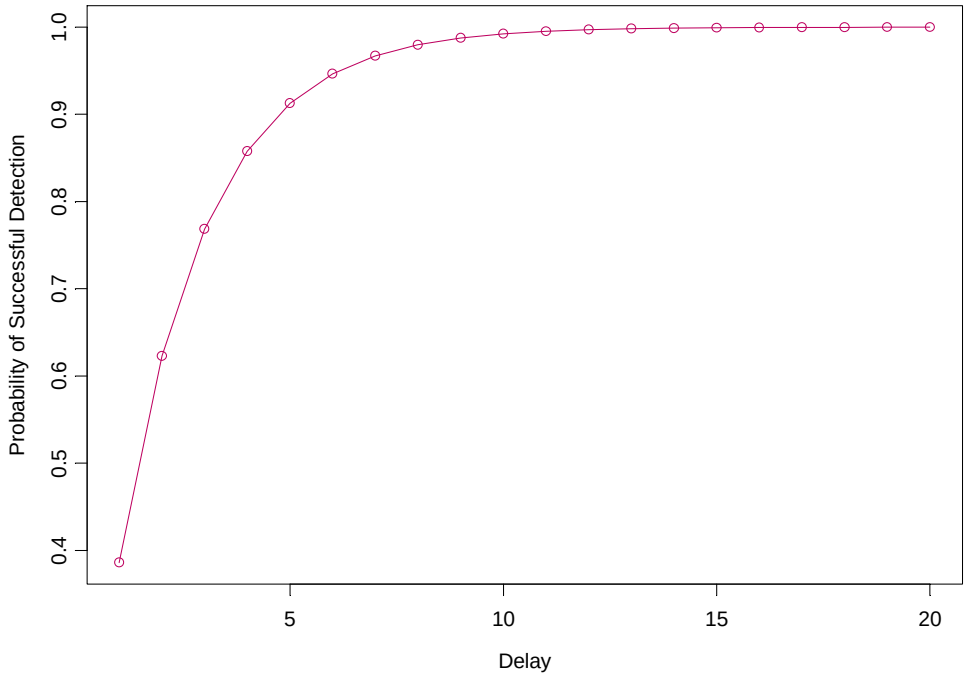


Figure 3.4: The visual projection of the PSD for the Shewhart scheme of the example in section 3.1

3.4 Predictive Value

If we have an alarm in our surveillance we have to choose the appropriate preventive actions. Fulfilling that purpose in practice, means knowing how much trust to put in an alarm. Additionally, it gives us information about which action would be appropriate. It is simpler for us, if the same action can be used whenever an alarm occurs. Thus, a constant predictive value with respect to the time is a good property.

For different methods, we have different false alarm distributions as a function of time. That fact leads to the principle that the proportion of false alarms compared with justified alarms at a specific time point will differ between the methods and so will the *trust* in an alarm. This criterion gives us a

quantitative idea of how strong an alarm is, as an indication of a change. Moreover, it tells us how probable a change in the past is, when we have an alarm and as a result it can inform us what will happen if no actions will be taken in the future. In order to express that *trust*, we use the measure of the *Predictive Value*.

Definition:
The Predictive Value is the probability that the process is out of control when an alarm is triggered and it is given by the following expressions:

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

3.5 Conditional Expected Delay

A measure of evaluation with respect to a true change in the vast literature is summarized by the average out-of-control run length $ARL_1 = E[t_A \mid \tau = 1]$.In this measure we make the assumption that the change occurred exactly the same time when the surveillance started. As we mentioned previously in the field of public health this assumption is unrealistic and useless. In public health surveillance the ability to detect a change depends on the time that the change occurred. So, we should take into account the possibilities of later changes.

That is a case which we are not going to use in this study but it is mentioned since it is referred in the vast literature of the evaluation of a method. Our measure is the *conditional expected delay* as a function of the time point *t*.

Definition:
The Conditional Expected Delay is the average delay time for a motivated alarm when the change occurs at a time point t and it is given by the following expression:

$$CED(t) = E[t_A - \tau \mid t_A \ge \tau = t]$$

3.6 Expected Delay

Since the case of $\tau=1$, as we saw before, is not the only case we are interested in, we can calculate the expected delay for other values of τ. In fact we are based on the distribution of τ, which is often geometrical, to consider the *expected delay.*

Definition:
The Expected Delay is a weighted average of the time between the change and the time that an alarm is triggered for motivated alarms and it is given by the following expression:

$$ED_{\tau} = \sum_{t=1}^{\infty} P(\tau = t)P(t_A \geq t)CED(t)$$

3.7 Sensitivity-Specificity

As mentioned above in a surveillance system we have to evaluate also its timeliness and the quality of the data collected. Important delays in detection may appear because of delays in reporting or confirming a diagnosis and data may be missing, wrong, misinterpreted etc. For that reason we use measures such as sensitivity and specificity.

In order to confirm our research a lot of times we do some further diagnostic tests in order to confirm if a disease is present or not. In this study we are not referring further in such measures but we present them for general knowledge reasons.

An example of these measures could be the study of Rolfhamre and Ekdahl(2006)[119] who used the sensitivity and the positive predictive value to compare the Poisson CUSUM (which we examine later in this study) and two other methods applied in different regions.

We have the following 2x2 table with the number of people for every possible combination, which will help us determining these two measures:

Table 3.2: The 2x2 table for the number of incidents according to the positive or negative result of the diagnostic test and the presence or absence of a disease

	Presence of Disease D^+	**Absence of Disease** D^-
Diagnostic Test Positive T^+	a	b
Diagnostic Test Negative T^-	c	d

Definition:
Sensitivity of a diagnostic process is the probability the diagnostic test to be positive given that someone has the disease. It is given by the following expression:

$$sensitivity = P(T^+ \mid D^+) = \frac{\#(T^+ \cap D^+)}{\#(D^+)}$$

Definition:
Specificity of a diagnostic process is the probability the diagnostic test to be negative given that someone does not have the disease. It is given by the following expression:

$$specificity = P(T^- \mid D^-) = \frac{\#(T^- \cap D^-)}{\#(D^-)}$$

Where our symbolism in these definitions represents quantities of the 2x2 table mentioned above:

Table 3.3: Interpretation of the symbols in the definitions of Sensitivity and Specificity

Symbolism in Definitions	***Interpretation in 2x2 Table***
$T^+ \cap D^+$	a
$T^- \cap D^-$	d
D^+	$a+c$
D^-	$b+d$

Chapter 4
Optimality Criteria

Optimality of different methods is a classical issue in surveillance. How they link to different measures of evaluation and how we make optimal a method through likelihood functions is the objective of the paper of Frisén(2009)[40]. A discussion on this study was presented by Knoth(2009)[69]. Further reading in optimality is given by Frisén(2010)[41] where the study of Shiryaev on the quickest detection problems is discussed.

At this point, we will use the measures of the previous section to formulate and discuss some criteria of optimality for surveillance.

4.1 Minimal Expected Delay

A general utility function was proposed by Shiryaev (1963)[128].In that case we are dealing with the expected delay of an alarm. He treated the case of constant intensity of a change where the gain of an alarm and the loss of a probable false alarm are a linear function of the value of the delay, $t_A - \tau$. This utility can be expressed as $U = E\{u(\tau, \mathrm{t_A})\}$, where:

$$u(\tau, \mathrm{t_A}) = \begin{cases} h(t_A - \tau), t_A < \tau \\ a_1(t_A - \tau) + \alpha_2, otherwise \end{cases}$$

The function $h(t_A - \tau)$ is usually a constant *k*, since we have a constant cost of triggering a false alarm independently of how early the false alarm is given. In this case we have:

$$U = bP(t_A < \tau) + \alpha_1 ED + a_2$$

Achieving a maximal utility corresponds to a *minimal expected delay from the change-point for a fixed probability of a false alarm*. This criterion is also known as the *Expected Delay criterion*. The full likelihood ratio method which we are going to examine in chapter 8 satisfies this criterion. The *ED criterion* seems to be a suitable optimality criterion in a public health setting because of its generality of including changes occurring at different time points.

4.2 Minimax optimality

This criterion concerns the *minimax of the expected delay after a change*. Despite the fact that several possible change times are considered (thus this criterion is related to the *ED criterion*), we use the *conditional expected delay* as it is stated in chapter 3. Instead of using an expected value, which requires a distribution of the time of change, we use the worst value of *CED(t)* avoiding at the same time any requirement of information about the distribution of τ. There is a lot of theoretical research based on this criterion.

Pollak(1985)[97] uses the worst value of τ to give an approximate solution to the criterion of *minimal expected delay*. He starts the procedure avoiding the properties which are dependent to the time of change τ.

Moustakides (1985)[91] uses a still more pessimistic criterion, since it is based on the worst possible circumstances. The worst possible case is considered, by using not only the worst value of the change time, but also the worst possible outcome of $X_{\tau-1}$ before the change occurs.

Ritov(1990)[108] considers a loss function which is not identical to that of Shiryaev(1963)[128] but depends on the time of change τ and the time of the alarm t_A. In this case we consider the worst possible distribution for each decision time s, $P(\tau = s+1 \mid \tau > s)$.With this assumption the CUSUM method minimizes the loss function.

Further studies for the *minimax optimal criterion* are given by Yashchin(1993)[171], Lai(1995)[76]&(1998)[77] and Lai & Shan(1999)[78].

4.3 Average Run Length

In this case optimality may be defined as *minimal* ARL_1 *for a fixed* ARL_0. The assumption that these two expectations make, is that there are equal distributions for all observations under each of the two alternatives. *ARL's* position in statistical quality control is a dominating one, since it is the most common used measure for an evaluation of a method. Therefore its usage as an optimality criterion in the industrial quality control and the other developing fields is widely known. However, its dominating position among other optimal criteria is doubted; especially when the field of interest is the public health.

We present some of the consequences of the usage of such a measure as an optimal criterion. Frisén(2003a)[37] shows that *there are values* c_s *such that a surveillance system with alarm at* $t_A = \min\left\{s : \sum_{t=1}^{s} X(t) > c_s\right\}$ *gives the minimal* ARL_1 *for a fixed value of* ARL_0.

Thus, a linear combination with equal weights to all observations fulfills this criterion of optimality (minimal ARL_1 for a fixed value of ARL_0) .On the other hand, methods with equal weights for old and recent observations are not appropriate(or at least rarely they are appropriate methods). So, this statement shows that this optimality criterion could be doubted. In the field of public health it is more probable not to have equal weights for our observations. Time is an important factor which can relate with other factors such as the spatial factor and so the case of the equal weights in practice is almost impossible.

Additionally, in the applications for which this optimal criterion is appropriate, the knowledge of the alarm statistic for each time of decision *s* is not enough. We should also have to determine the alarm limit c_s for this statistic for each *s*.

Frisén proceeded in the construction of a two-point method which fulfills the *criterion of minimal* ARL_1 *for a fixed* ARL_0.The method has the alarm limits:

$$c_1 = L, c_i = \infty, i = 1,2,...,k-1 \text{ and } c_k = -\infty$$

where $k = \dfrac{ARL_0 - \Phi(\text{-L})}{\Phi(\text{L})}$ for the standard normal distribution function Φ.

From the above we have that *L* is restricted to those values and as a result *k* is an integer. It was proved at the same paper that *this two point method*

fulfills the criterion of minimal ARL_1 *by having* ARL_1 *arbitrarily close to the minimal value of one, for a fixed value of* ARL_0.

This method, though, has very bad properties. It depends only on the first observation regardless if an alarm is given at time 1 or at another time *k*.

For that reason Frisén used the LCUSUM method with a time of an alarm at $t_A = \min\left\{s : \sum_{t=1}^{s} X(t) > L + \frac{s\mu}{2}\right\}$ (it is described in the paper's appendix) which *minimizes the* ARL_1 *for a fixed false alarm probability.*

Pollak and Siegmund (1985)[99] pointed out that the maximal value of CED(t) is equal to *CED(1)* for many methods and with a minimax perspective this can be a motivation for the use of ARL_1 since $CED(1) = ARL_1 - 1$.However this argument is not relevant for all methods(*CED(1)* is not the maximal value for the EWMA method as presented by Frisén and Sonesson(2002)[47]).

Thus we see that this optimal criterion has many disadvantages. Frisén pointed these flaws with the three statements mentioned above. Even if this criterion is used in a method there is a strong indication that it is not reliable. We see that methods useless in practice are *ARL* optimal and that this criterion cannot be used for all the methods.

Hence, we should use this criterion of optimality only with care and we have to be cautious about the properties of each method we use. The Average Run Length can give us some information as descriptive measure of evaluation but as an optimal criterion has many flaws and disadvantages.

PART I

Detection of Increased Rates of Incidence in a Poisson Process


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 5
Using the time between events to study the Poisson process

When we are referred to a Poisson process an increased rate of incidence corresponds to an increased intensity of the Poisson process. Thus, the possibility of detecting an increased intensity depends on the method we use to monitor the process. The way we observe our process is important too. Different techniques of handling a Poisson procedure are mentioned below.

Here, we observe the time intervals between two adverse health events of our procedure. With the term “adverse health events” we mean the presence of a disease. Thus, we can measure these time intervals with the following two ways:

a) In the case of the *continuous time*, the intervals are distributed exponentially.
b) In the case of a *discrete time scale*, we count the number of acceptable events between adverse events.

Both these ways include no loss of information about the process. Increased intensity or an increased rate of a disease can be interpreted with two ways respectively:

a) In the case of the *continuous time*, increased rate means shorter intervals between two adverse events.
b) In the case of the *discrete time*, increased rate means smaller number of acceptable events.

In the same way we have different designs of methods according to the case we are interested in:

a) In the case of the continuous time between two adverse events, which is distributed exponentially as we mentioned above, we are interested in the design of methods like the **exponential CUSUM** and the **EWMA(exponentially weighted moving average)**.

b) In the case of the discrete time there are methods such as the **Sets method** and the **Cuscore method.**

In the field of the surveillance in public health, the first case of the continuous time is very rare which is why we are devoted, in this study, to the second case of the discrete time.

5.1 The Sets Method

The original sets method was proposed by Rina Chen(1978)[15] for the surveillance of congenital malformations. Since then, there have been a lot of improvements for this method. We give a general description of this method.

The sets method focuses on the lengths of the intervals between events with the presence of a disease. These lengths are measured with the number of healthy cases between two events with the disease we are interested in. A first assumption we make is that these lengths are distributed geometrically. An alarm is triggered in our system when *n* consecutive intervals are shorter than some threshold value. That is our system's function which we are going to present by simulations in this section.

This method is really simple, easy to understand and can be applied in many different cases for a variety of diseases. However, this method presents some disadvantages as far as we are concerned about its optimality. The sets method is not constructed from likelihood ratios, which is an optimal method (we are going to describe this method later) and considers for the alarm decision only the data which are based on the last *n* consecutive intervals. So, we might have as a result a huge amount of loss of information. In this chapter we are going to describe the method through some applications and present the improvements that have been made by some researchers.

We are studying the Rina Chen's application of the congenital malformations. In this example we are interested in the intervals between two newborns with a congenital malformation. The length of these intervals is measured by the number of healthy newborns. The calculations are simple and this simplicity makes the method easy to be used. We are presenting the cases of small and large scale systems (one hospital and several hospitals respectively).

The normal births between two births with the specific malformation monitored, is defined as a *set*. Each time the epidemiologists diagnose a new case of disease (the disease in our example is the congenital malformation of newborns), they move on to an analysis of the last *n* intervals (sets).These intervals or sets constitute our sequence of interest.

The size of a *set* (the malformed newborns are excluded) is a geometric variable. We have an alarm, if a sequence of sets appears in a way that each set is below a fixed size. In practice, this sequence is more frequent to happen for the case when the malformed newborns have an increased rate and less frequent to happen when the rate of malformed babies is normal. The hypothesis we test is expressed as follows:

H_0: *The rate of malformed newborns is normal.*

H_1: *The rate of malformed newborns is increased.*

5.1.1 Probabilities and Expectations

We are interested in estimating the number of infants expected to be born before an alarm is triggered. Thus, we study the number of normal births *X* until we have a success where in our example the "success" is a newborn with malformation.

The random variable X_i is the size of the i^{th} set and it is assumed to follow a geometric distribution. We remember that if ***X~geom(p)*** then:

$$P(X = x) = p * (1-p)^x$$
$$E(X) = \frac{1-p}{p}$$
$$Var(X) = \frac{1-p}{p^2}$$

We use the following notation:

Table 5.1: The terms and their explanation we use for the Sets method

Terms	*Explanation*
p_0	The normal rate of the malformed newborns. Often an estimation for p_0 is given by the rate of malformations over several past years.
p_1	The increased rate of the malformed newborns. We also assume that: $p_1 = \gamma * p_0$.
a_1	The number of the expected births in a sequence initiated after the increase which signals an alarm.
t_1	The expected time duration in which the a_1 infants are born.
P_i	The probability that, under the hypothesis H_i, a given sequence signals an alarm.
r	The expected number of false alarms during a time interval.
b	The expected number of births during a time interval.

Thus, the hypothesis test we mentioned previously can be expressed as follows:

$$H_0 : p = p_0$$

$$H_1 : p = p_1 \ or \ \ p = \gamma * p_0$$

❖ ***The case under the null hypothesis*** H_0

We suppose we are in the case where the null hypothesis stands. Then, under H_0, we have the *expected size* of the set *i:*

$$E(X_i) = \frac{1-p_0}{p_0} = c_0 \quad (5.1)$$

Then the *probability a size of a set to be smaller than* $E(X_i)$ is:

$$P(X_i < E(X_i) = c_0) = \sum_{X=0}^{c-1} p_0(1-p_0)^X \Leftrightarrow \quad (5.2)$$
$$P(X_i < E(X_i) = c_0) = 1-(1-p_0)^{c_0} = 1-e^{-1} = 0.632$$

(Since we have for a very large $\frac{1}{p_0}$ that $(1-p_0)^{kc} = e^{-k}$ for any k>0).

Generally it is stated that:

$$P(X_i < kc_0) = 1-(1-p_0)^{kc_0} = 1-e^{-k} \quad (5.3)$$

The moment generating function of X_i given $X_i < k * c_0$ is:

$$M(t) = \frac{1}{1-(1-p_0)^{kc_0}} \sum_{X=0}^{kc-1} p_0(1-p_0)^X e^{tx} = \frac{p_0}{1-(1-p_0)^{kc_0}} \frac{1-(1-p_0)^{kc_0} e^{tkc_0}}{1-(1-p_0)e^t}$$

thus,

$$E(X_i \mid X_i < kc_0) = M'(0) = c_0\left\{1-\frac{k}{e^k-1}\right\}, i = 1,2,...$$

❖ ***The case of the alternative hypothesis*** H_1

With the same way we act in the case when we are under H_1:

$$c_1 = E(X_i) = \frac{1-p_1}{p_1} = \frac{1-\gamma * p_0}{\gamma * p_0} = \frac{1}{\gamma} * \frac{1-p_0}{p_0} = \frac{1}{\gamma} * c_0$$

From *(5.3)* we have:

$$P(X_i < kc_0) = 1-(1-p_1)^{kc_0} = 1-(1-p_1)^{k\gamma c_1} = 1-e^{-k\gamma} \quad (5.4)$$

Similarly we calculate:

$$E(X_i \mid X_i < kc_0) = c_1\left\{1-\frac{k\gamma}{e^{k\gamma}-1}\right\} \quad (5.5)$$

5.1.2 A System for Surveillance in a Single Hospital

We are now interested in *n* consecutive sets in a single hospital. The rule of an alarm is based on the size of each set we study. That fact actually means that an alarm will be triggered by a sequence of sets, if each size of these sets is below a certain size.

If, in our simulation, X_i for *i =1,2,3,…,n* is the size of each set, then in *n* consecutive sets, which is our sequence, we will have a $X_{\max}$.With this term we define the largest size of the sets in our sequence.

Then we shall have the **probability that a sequence will trigger an alarm:**

$$P(X_{\max} \le kc_0) \quad (5.6)$$

The probability is defined in our hypothesis test as follows:

$$H_0 : P_0 = P(X_{\max} \le kc_0) = (1-e^{-k})^n \quad (5.7)$$

$$H_1 : P_1 = P(X_{\max} \le kc_0) = (1-e^{-k\gamma})^n \quad (5.8)$$

In the case of the null hypothesis we want the probability to be significantly low so that the probability of a false alarm will be low and in the alternative case we want the probability P_1 to be high so that an alarm is

justified and will be triggered when the rate of the malformed newborns is increasing.

Supposing that $(1-e^{-k\gamma})=0.99$ (so that P_1 will be large) we should have:

$$k\gamma = -\ln 0.01 \Leftrightarrow k = \frac{4.61}{\gamma} \quad (5.9)$$

And so for the null hypothesis we should have the following equation for the size of a sequence of sets:

$$n \approx \frac{\ln P_0}{\ln(1-e^{-k})} \quad (5.10)$$

It is reasonable to assume that under the null hypothesis the number of expected newborns having the particular malformation is:

$$b * p_0 \quad (5.11)$$

Also, we can assume the number of sequences (of size *n*) among *b* newborns to be:

$$b * p_0 - (n-1) \quad (5.12)$$

So, the average number of false alarms is given by:

$$r = \{b * p_0 - (n-1)\} * P_0 \quad (5.13)$$

Reasonably, the probability P_0 is given by the following equation:

$$P_0 = \frac{r}{\{b * p_0 - (n-1)\}} \quad (5.14)$$

We also can define from *(5.5)* the number of the expected births in a sequence which signals an alarm after an increase:

$$a_1 = n * c_1 \left\{ 1 - \frac{k\gamma}{e^{k\gamma} - 1} \right\} \quad (5.15)$$

Simulation

For our example of congenital malformations of newborns, we are interested to make a surveillance system in a single hospital. Suppose now that we calculated the number of births in a particular hospital to be about 400 births per month. We also suppose that we have one false alarm in 20 years. Our surveillance system realizes as an important change to signal an alarm, an increased rate seven times the normal rate of congenital malformations in newborns. From these data we have:

$$\gamma = 7$$
$$r = 1$$
$$b = 400 * 12 * 20 = 96000$$
$$k = \frac{4.61}{7} = 0.658$$

We have the following ten fixed "normal" rates $p_0 = \frac{1}{10000}, \frac{2}{10000}, ..., \frac{10}{10000}$

For these fixed "normal" rates we have the following table:

Table 5.2: The table with the results from the simulation of the Sets method for different baseline rates

$p_0 \times 10^{-4}$	P_0	c_0	n	a_1	t_1
1	0.13157895	9999	3	4088	10.220697
2	0.06172840	4999	4	2725	6.813117
3	0.03875969	3332	4	1817	4.541623
4	0.02906977	2499	5	1703	4.257346
5	0.02272727	1999	5	1362	3.405536
6	0.01865672	1666	5	1135	2.837663

7	0.01607717	1428	6	1167	2.918447
8	0.01392758	1249	6	1021	2.553385
9	0.01228501	1110	6	908	2.269449
10	0.01098901	999	6	817	2.042299

To calculate the results of the table above we worked in S-Plus. The code is given to *Appendix A.1*. For this example we could make several conclusions depending in the kind of results we are interested in:

- For a normal rate of congenital malformations of $p_0 = \frac{1}{10000}$ (one malformation per ten thousand births), we have that the expected size of a set, under the hypothesis that the rate is "normal", is 9999 births.

- For the same case, we have that an alarm should be signaled after 3 consecutive sets, each smaller than $0.658 * 9999 = 6579$ births.

- Also for the same case, the probability having a false alarm after 3 consecutive sets is 0.132 or 13,2% and the probability that an increased rate seven times the normal rate would be detected after a sequence of 3 sets, is $P_1 = P(X_{\max} \le kc_0) = (1 - e^{-k\gamma})^n = 0.971$ or 97,1%.

- The expected number of newborns included in the above sequence which will trigger an alarm after an increase, is 4088.

- The expected time where we will have an alarm after the increase in the rate is 10,2 months.

We also can make some general comments of vital importance for this method:

- The probability having an alarm when the normal rate still stands is small for a large (fixed) 'normal' rate. Thus the rarer the disease, the more probable having a false alarm.

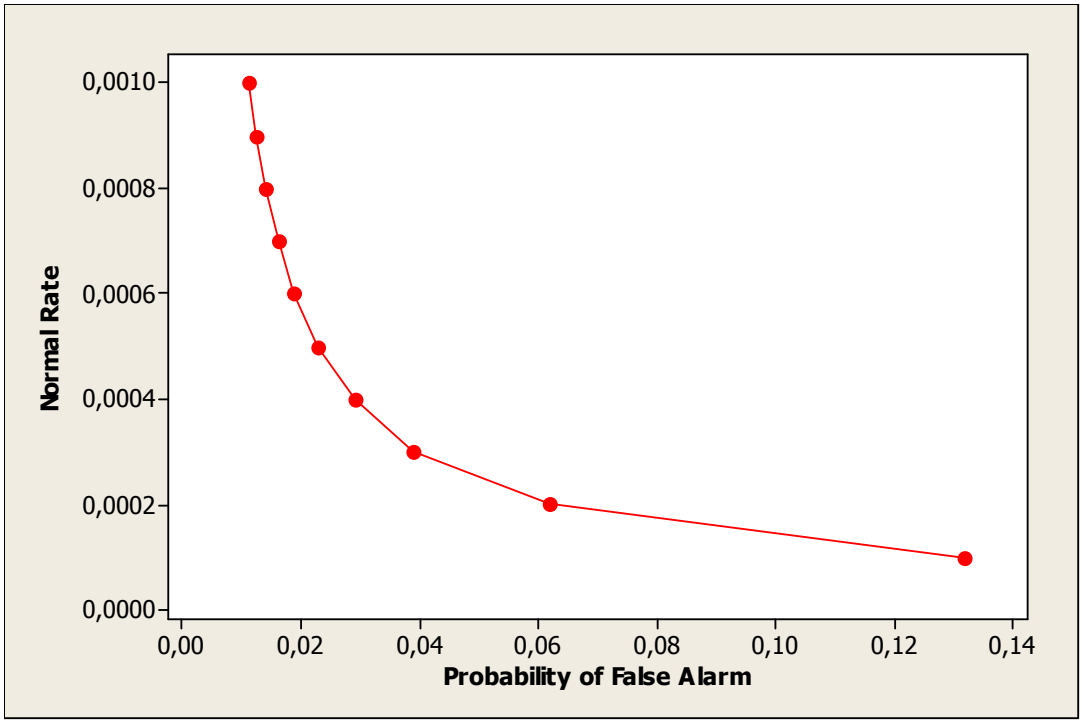


Figure 5.1: The PFA according to the baseline rate of a disease(newborn's malformation) for a single hospital

➢ Using the same logic, we need fewer infants and subsequently less time to realize the change in the rate of a frequent disease compared with a rare disease.

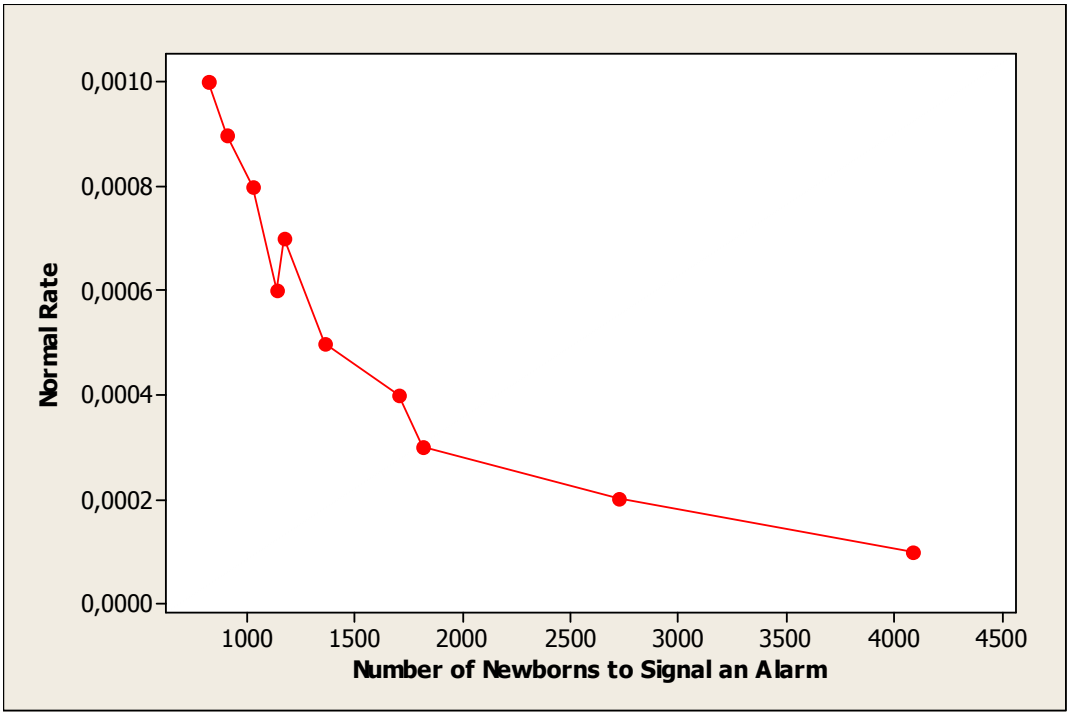


Figure 5.2: The number of newborns until the signal of an alarm according to the baseline rate of a disease (newborn's malformation) for a single hospital

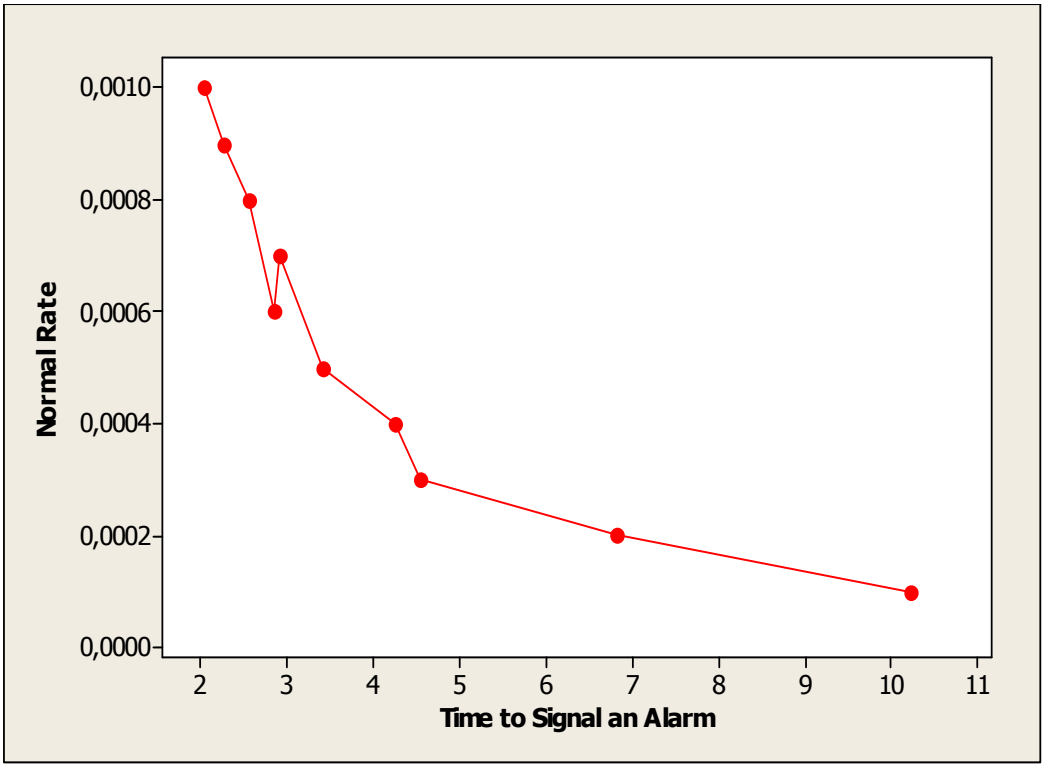


Figure 5.3: The time until the signal of an alarm according to the baseline rate of a disease (newborn's malformation) for a single hospital

➢ From the above we realize that the sets method proposed in this section is better for detecting changes in the rates of diseases with a large base-line rate. The probability of a false alarm is significantly large for small "normal" rates and simultaneously our system realizes far more sooner the change when we have large "normal" rates. Subsequently, our method is much better when the disease is frequently appeared in the population of interest.

Especially, when we are interested for rare diseases and an alarm is triggered in our system, we should not take for granted that this is a true alarm. It is better to assume that this alarm is more than a warning. Then the epidemiologists should investigate the case in order to find proof leading to an alarm.

For this reason, we should take a larger *r(number of false alarms)*into the single hospital system in order to have a shorter time delay between the change and triggering a warning. As a result, we should have a true alarm in a shorter time interval but on the other hand this fact leads to more frequent false alarms. For our example we should

have the following table for a number of false alarms $r = 5$ in 20 years. The code is given in *Appendix A.2*.

Table 5.3: The table from the simulation of the Sets method for different baseline rate and 5 number of false alarms in 20 years(r=5)

$p_0 \times 10^{-4}$	P_0	c_0	n	a_1	t_1
1	0,657895	9999	1	1363	3,406899
2	0,274725	4999	2	1362	3,406558
3	0,179856	3332	2	908	2,270812
4	0,137363	2499	3	1022	2,554408
5	0,108696	1999	3	817	2,043322
6	0,089928	1666	3	681	1,702598
7	0,077882	1428	4	778	1,945631
8	0,067751	1249	4	681	1,702257
9	0,059952	1110	4	605	1,512966
10	0,053763	999	4	545	1,361533

- For a normal rate of congenital malformations of $p_0 = \frac{1}{10000}$ (one CM per ten thousand births), we have that an alarm should be signaled after 1 set smaller than 6579 births instead of 3 consecutive sets when we have one false alarm in 20 years.

- The expected number of newborns which will be born until our system realizes the increase and trigger an alarm is 1363 instead of 4088 newborns in the case when we had one false alarm in 20 years.

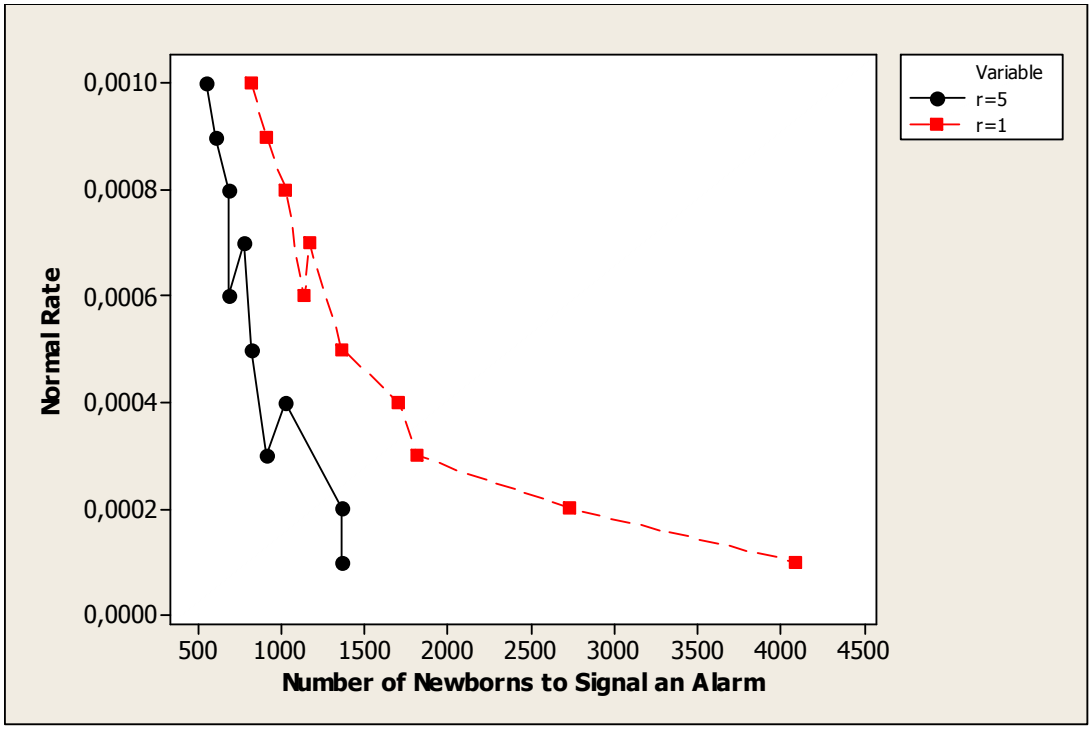


Figure 5.4: Comparison of the number of newborns until the signal of an alarm according to the baseline rate of a disease (newborn's malformation), for one and five false alarms in 20 years

➢ Additionally the time duration until our system realizes the increase in the rate and triggers an alarm is much less, compared with the case of the one false alarm. Specifically, our system in the case of 5 false alarms in 20 years will have realized an increase in a rare disease in 3 to 4 months compared with the previous case where an increase of a rare disease is detected in 10 to 11 months.

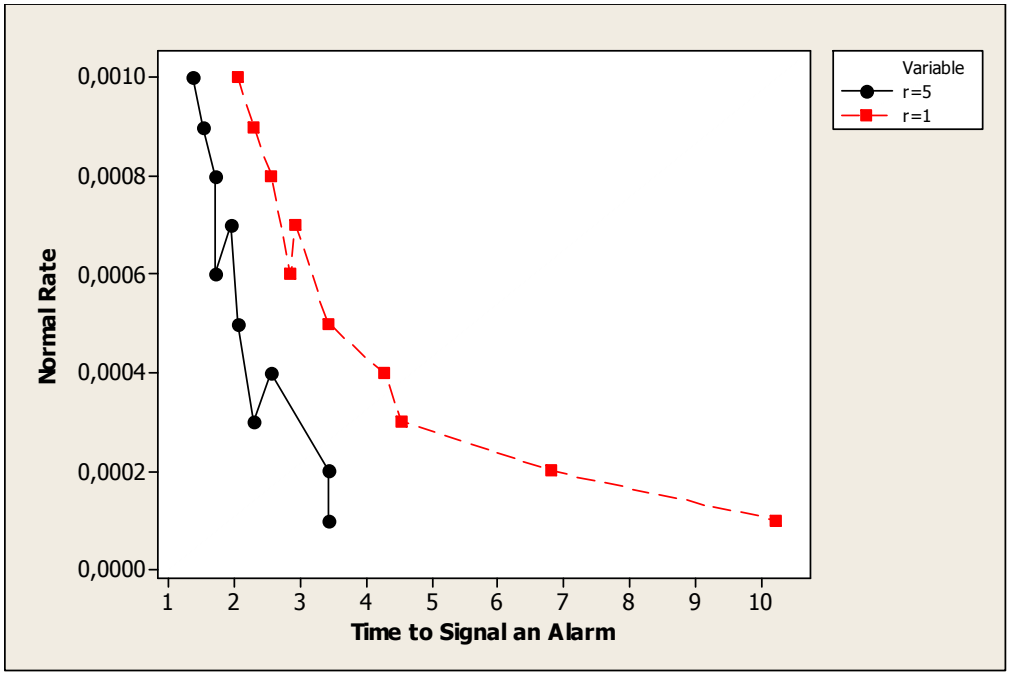


Figure 5.5: Comparison of the time until the signal of an alarm according to the baseline rate of a disease (newborn's malformation), for one and five false alarms in 20 years

➢ As we mentioned above, this increase in the number of false alarm has a side effect. That is the large probability of having a false alarm. In our case this probability is 65,7%(number of false alarms in 20 years is 5) while in the previous case is 13,2%(number of false alarms in 20 years is 1) for the smallest 'normal' rate .Additionally, we have that the probability an increased rate seven times the normal rate would be detected after a sequence of 1 set, is $P_1 = P(X_{\max} \le kc_0) = (1 - e^{-k\gamma})^n = 0.99$ or 99%.

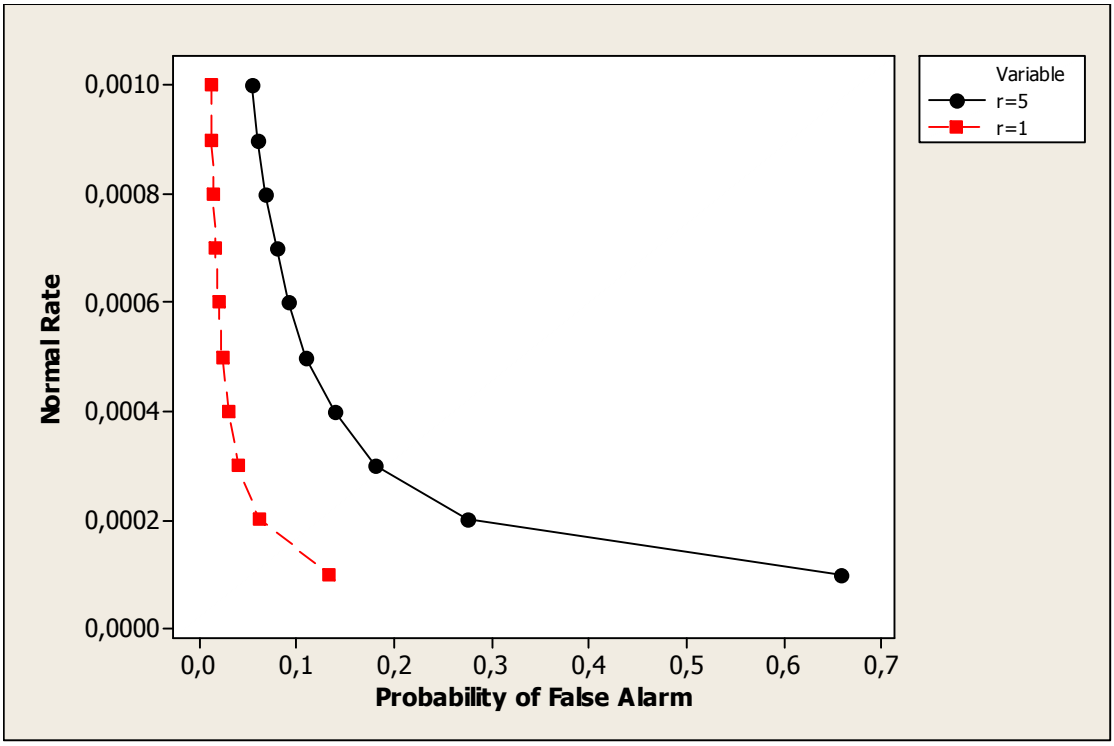


Figure 5.6: Comparison of the PFA according to the baseline rate of a disease (newborn's malformation), for one and five false alarms in 20 years

5.1.2.1 A Second Approach to find the n

In the approach mentioned above, we calculated the number of sets *n* using a cross-checking technique for *(5.10)* and *(5.14)* equations. Here, we are going to use the same cross-checking technique but for the *(5.7)* and *(5.14)* equations. In this section we act based on the probability of a false alarm. We simulate some tests for a case and we choose the appropriate *n* based on the similarity of the values. In our example we have a number of false alarms $r = 1$ and we have $k = 0.658$ for $\gamma = 7$ .Lets take the example of the baseline rate of $\pi_0 = 5 * 10^{-4}$.For the simulations, we used S-plus and the code is given in the *Appendix A.3*.

Then we should have the following table:

Table 5.4: The decision of the appropriate *n* for a baseline rate of 5/10000 and one false alarm in 20 years

n	$P_0 = \frac{r}{\{b * p_0 - (n-1)\}}$	$P_0 = (1-e^{-k})^n$
1	0,0208333	0,482114
2	0,0212766	0,232434
3	0,0217391	0,112060
4	0,0222222	0,054025
5	0,0227273	0,026046
6	0,0232558	0,012557

From the table above we see that the minimum difference between the two equations exists for ***n=5***.So we should take *n=5* for the case of $\pi_0 = 5 * 10^{-4}$.Making comparisons for all the cases of the normal rate in our example we should have the following table:

Table 5.5: Comparing the decision of *n* according to the process we chose to use for different baseline rates.

$p_0 \times 10^{-4}$	***n from equations (5.10) and (5.14)***	***n from equations (5.7) and (5.14)***
1	3	3
2	4	4
3	4	5
4	5	5
5	5	5
6	5	6
7	6	6
8	6	6
9	6	6
10	6	6

We notice that there is a difference in two cases ($\pi_0 = 3*10^{-4}$ and $\pi_0 = 6*10^{-4}$). In this approach the number of sets for these two cases is larger than the previous one and so the number of newborns included in a sequence which signals an alarm after an increase and the time expected our system to realize the change, will be larger. Specifically, the number of newborns will be 2271 instead of 1817 births and 1362 instead of 1135 births respectively for each case.

The advantage of this approach is that the probability of a false alarm for these two cases is smaller (2.6% instead of 5.4% from *(5)*).Also the probability that an increased rate seven times the baseline rate would be detected after a sequence of 5 sets is 0.95.**This method is better for smaller increases to be detected with some reasonable probability but after a larger number of diagnoses.**

5.1.2.2 A Third Approach to find the n

At this point we could use another approach to calculate the number of consecutive sets from which the increase will be detected. In the approach mentioned above, we calculated the number of sets *n* using a cross-checking technique for *(5.10)* and *(5.14)* equations and for *(5.7)* and *(5.14)* equations.

Previously, we used a fixed value for *γ*. Here, we are searching for the appropriate *γ-value* which will lead us with more precise calculations to the appropriate *n*. In this section we are going to figure out the *n*, using a criterion which maximizes the probability of detecting a change in the baseline rate.

We take the symbolism for time reasons of $M = p_0 * b$.With *M* we denote the expected number of cases between false alarms. Thus, the value of *M* is chosen prior to our system's beginning. Using this symbolism the *(5.14)* equation is calculated as follows:

$$P_0(n) = \frac{r}{M - n + 1} \quad (5.16)$$

A way to express the exact value of M is by the following equation[1]:

$$M = \frac{1 - P_0(n)}{[1 - P_0(1)] * P_0(n)} \quad (5.17)$$

So for equation *(5.7)* we have:

$$P_0(n) = (1 - e^{-k})^n \Rightarrow P_0(1) = 1 - e^{-k} \quad (5.18)$$

The equation *(5.17)* for M now should be:

$$M = \frac{1 - (1 - e^{-k})^n}{e^{-k} * (1 - e^{-k})^n} \quad (5.19)$$

As we said above the value of M is chosen prior to the system's start. Thus, for a given n, we can find the k-value. We also have for γ from *(5.8)* that:

$$\gamma(n) = \frac{-\ln\{1 - P_1(n)^{1/n}\}}{\mathrm{k(n)}} \quad (5.20)$$

From these equations we have that if n increases then k will increase (γ will decrease). But, if the number of sets is increasing then the probability to detect a change is decreasing. **Hence, there is a maximum value for that probability** P_1.

The main idea for this method is that for a given γ, we can determine the value of n which maximizes that probability. This procedure is feasible as $\gamma(n)$ is a monotone decreasing function of n.

The rule to determine the n is based on the fact that **the efficiency of a single analysis of *n* intervals does not increase much relative to an analysis involving *n+1* intervals.** The actual value of n used for the alarm detection is determined as the smallest value for which:

$$\gamma(\mathrm{n}) - \gamma(\mathrm{n}+1) \leq 1$$

[1] This is shown in Kenett & Pollak(1983)[67]

Note: A false assumption is common to be made at this point. That is the one false alarm expected in *M* cases. Then we should have for $P_0(n)$ the following equation from *(5.16)*:

$$P_0(n) = \frac{1}{M - n + 1}$$

Solving the equation *(5.16)* for *M* and for one false alarm we have the approximation equation for *M*. This equation would be correct **only if** we have in our research **a finite number of gaps (cases)**.In real life this is not true. When we want to make tests with that system, in practice means continuous analyses for infinite cases. Thus, this approximation has a significant difference from the exact value. Taking the ratio of the approximation equation for *M (from 5.16)* and the equation *(5.17)* to prove that fact, we have that[2]:

$$M_{appr} = \frac{1}{P_0(n)} + n - 1 \text{ and } M = \frac{1 - P_0(n)}{[1 - P_0(1)] * P_0(n)}$$

Eventually from these two equations we should have the following ratio:

$$\frac{M}{M_{appr}} = \frac{1 - (1 - e^{-k})^n}{e^{-k} * \{(n-1) * (1 - e^{-k})^n + 1\}}$$

Taking for example the case of *n=5* and *k=0.76* we should have

$$\frac{M}{M_{appr}} = \frac{0.957}{0.547} = 1.75$$

Interpreting this result, we can say that **using the specified values of *n* and *k*, the actual time interval between false alarms is 1.75 times the ones assumed with the approximation of M_{appr} .That is a substantial difference**

[2] The exact equation is shown in Kenett & Pollak(1983)[67]

between the approximate and the exact function. For these reasons we use the exact equation.

Simulation

For the example we used in the previous section, we have that *b*=96000.Suppose we are interested for a disease with a normal rate of $p_0 = \frac{5}{10000}$.Then the prior value of *M* (the expected number of cases within which one false alarm is expected) would be $M = p_0 * b = 48$.We also suppose that the probability to detect the hypothetical change in the ratio is 0.95. For this case we should have from our code[3] the following table:

Table 5.6: Choosing the appropriate *n* using the gamma rule.

n	***M***	***kappa***	***gamma***
1	48,1208	0,021	142,654
2	47,7920	0,169	21,752
3	48,0619	0,368	11,080
4	48,0668	0,569	7,668
5	48,0542	0,759	6,041
6	48,0219	0,936	5,092
7	47,9463	1,101	4,468
8	48,0081	1,254	4,029

For the table above we have the smallest *n* for which the rule for the gammas holds. That is ***n=5***.Comparing with the previous simpler method we see that the number of sets needed to detect the change is five too. The problem in that case was that we included the number of false alarms. In this section we do not have to make any arbitrary assumptions and we are based on a reliable rule which leads us to more precise and absolute conclusions.

All of these approaches are presented by Rina Chen's studies. Especially for the case of searching the appropriate *n*, the last approach seems to be the

[3] The code we used in S-plus is given in *Appendix A.4*

best since it does not take into account assumptions that may lead to unreliable conclusions. To be exact in our conclusions, this last method is better to detect small changes in the rate of a disease. That is the reason why it is preferred for the case of rare diseases. The first two approaches are based on the approximate equation for the expected number of diagnoses in a given period time. The last approach is better since it relies on the exact equation.

5.1.3 A System for Surveillance in Several Hospitals

In this section we are dealing with the increase in the rate of a disease in a larger area. For this reason we use all the available information from several hospitals. Our purpose is to detect increases which occur at the same time in an area of interest. The data we use in this system are the same with those recorded for the system of a single hospital.

What differ from the previous section are the criteria according to which an alarm would be triggered. In this large scale system we are proceeding in an analysis of the data at the end of a constant interval of time. In this constant interval we record the number of the sets and their size. In this system we have an alarm if each of the *m* completed sets is smaller than a value $k_m c_0$. Then we correlate the value of k_m with the probability of an alarm under the null hypothesis. This probability of false alarm given that at least one set is completed, is denoted by q_0.So, we have the following equations for k_m and q_0:

$$q_0 = P(X_{\max} < k_m c_0 \mid M = m) \text{ or } q_0 = (1 - e^{-km})^m \quad (5.21)$$

and

$$k_m = -\ln(1 - q_0^{1/m}) \quad (5.22)$$

The probability of false alarm is denoted by P_0 as we saw in the previous case of the single hospital. We denote with *y* the years for which the analysis takes place with a frequency of *d*-times per year. Then we should have the following equation:

$$P_0 = \frac{1}{y * d} \quad (5.23)$$

We make the assumption that the average number of false alarms is one (here we have an *average* number since we have several hospitals).

We mentioned above that m is the number of the completed sets. We symbolize with M the number of sets terminated within a certain period. Then the probability P_0 should be:

$$\begin{aligned} P_0 &= \sum_{m=1}^{\infty} P(X_{\max} < k_m c_0, M = m) = \\ &= \sum_{m=1}^{\infty} P(X_{\max} < k_m c_0 \mid M = m) * P(M = m) = \\ &= q_0 * \sum_{m=1}^{\infty} P(M = m) = \\ &= q_0 * \{1 - P(M = 0)\} \end{aligned} \quad (5.24)$$

With the term $P(M = 0)$ we denote the probability that none of the sets is completed during the surveillance in our time-interval. If we do not have a single set completed, in other words means that we have not diagnosed a single case with the disease monitored. Therefore, under the null hypothesis H_0 and for the N cases we checked for our disease of interest, we have:

$$P(M = 0) = (1 - \pi_0)^{\mathrm{N}} \quad (5.25)$$

So from the *(5.24)* and *(5.25)* equations we have the following:

$$P_0 = q_0 * \{1 - (1 - \pi_0)^{\mathrm{N}}\} \quad (5.26)$$

and

$$q_0 = \frac{P_0}{1 - (1 - \pi_0)^{\mathrm{N}}} \quad (5.27)$$

Simulation

In order to proceed in a simulation, suppose we have to apply a surveillance system in an area where a number of hospitals exists. We are interested in a type of cancer. For the purpose of our surveillance, we proceed in an analysis every 3 months of the 12.000 hospitalized individuals with an average of one false alarm in 40 years. For the data above we have:

$$y = 40$$

$$d = 4$$

Thus from *(5.23)* we have:

$$P_0 = \frac{1}{y * d} = 0.00625$$

And so we should have the following table[4]:

Table 5.7: The table for the simulation of the Sets method for different baseline rates for surveillance in several hospitals

p_0	c_0	q_0	***Kappa for m=5***	***Kappa**** c_0
0,0001	9999,00	0,0089436	0,493176	4931
0,0002	4999,00	0,0068734	0,460999	2304
0,0003	3332,33	0,0064255	0,453189	1510
0,0004	2499,00	0,0063018	0,450968	1127
0,0005	1999,00	0,0062655	0,450310	900
0,0006	1665,67	0,0062547	0,450113	750
0,0007	1427,57	0,0062514	0,450054	642
0,0008	1249,00	0,0062504	0,450036	562
0,0009	1110,11	0,0062501	0,450031	500
0,0010	999,00	0,0062500	0,450029	450

So, we have the following conclusions:

[4] The code we used in S-plus is given in *Appendix A.5*

➢ In the following graph we have a visual image of how the probability of a false alarm, given that at least one set is completed, works according to the baseline rate (normal rate) of a disease.

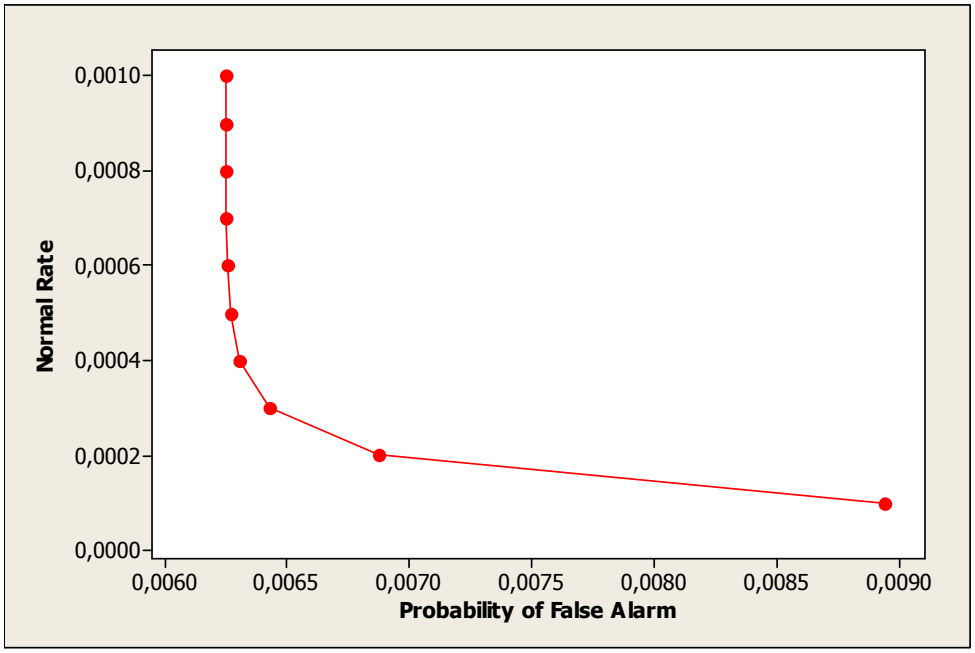


Figure 5.7: The PFA according to the baseline rate of a disease (newborn's malformation) for several hospitals

➢ Additionally, interpreting the table above, we have that if five sets have terminated within a three month interval, then we should have an alarm if each of the sets do not exceed a limit. This limit is given in the last column $(Kappa^{*}c_0)$.For a visual illustration of the limits of an alarm according to the baseline rate check the following graph:

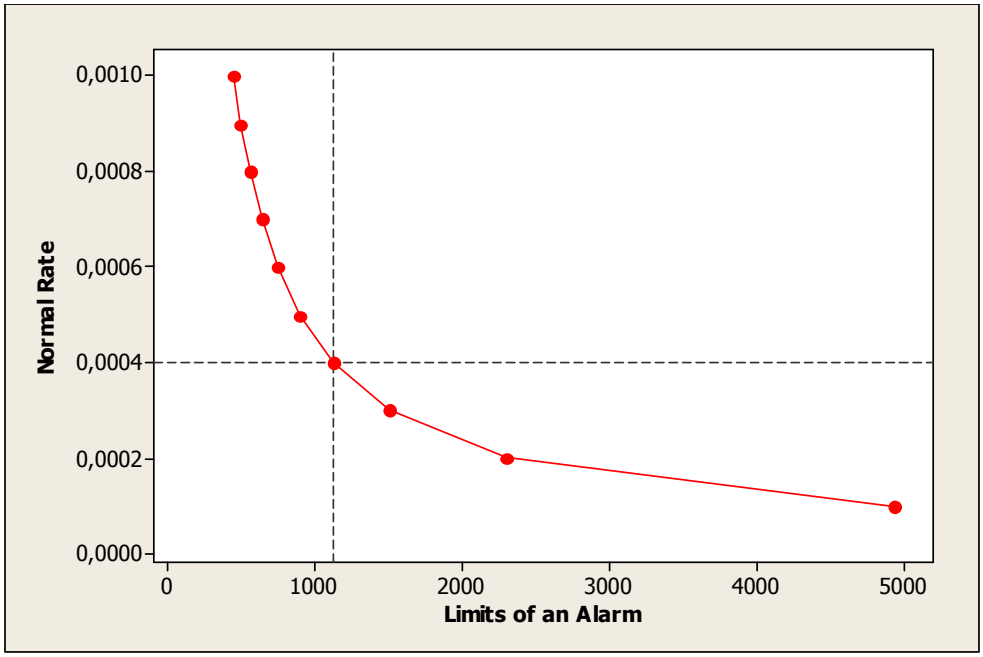


Figure 5.8: The limits for which we have an alarm if each of the sets do not exceed them (case of several hospitals)

- For example, for the case of the normal rate of 4 incidents-'events' per 10000 individuals, we should have the limit of 1127 cases. The proper interpretation would be that we have an alarm if each of the five sets does not exceed the value of 1127.

Note: If we would want to find the probability a gamma-time (the normal rate) increase to be detected, then our interest should be the probability under the alternative hypothesis. That probability should be:

$$P_1 = \sum_{m=1}^{\infty} (1 - e^{-k\gamma})^m P(M = m) \quad (5.28)$$

Further discussion on these matters can be shown in the study of Gallus et al.(1986)[50] in which the optimality of the sets method was approached and for which R.Chen was based to analyze the last procedure of finding *n*. Additionally other studies for this subject can be the one of G.Gallus et al.(1991)[51] in which from the main idea of the sets method a Poisson approximation is made to a Negative Binomial process and the one of Chen et al.(1997)[21]. The study of H.Arnkelsdottir(1995)[5] is really good and interesting since it presents some

measures (such as the Probability of Successful Detection and the Predictive Value) in order to evaluate the sets method.

5.2 The Cuscore Method

As mentioned above, the sets method's main disadvantage is the loss of information since that technique is based on the last *n* consecutive intervals. In order to solve this problem, a modification is required in a way that all observations until the last alarm are taken into account.

This new method is called Cuscore method and it is based on a score. Values of 1 or -1 are given to each interval between adverse events depending on whether or not it is longer than a threshold value. The statistic, which is based the 'alarm' event, is formed from the cumulative score.

With this technique, we avoid any loss of information which might be caused by the use of only *n* consecutive sets. Despite the fact that this method solves this problem (takes into account all the available sets), it has an important disadvantage. That is the type of reporting observations. When we intend to match the observations 1 and -1 with the real data, we have as a side effect a direct loss of information owing to the dichotomization of data which leads to a suboptimal method.

To make our procedure an optimal one, we are going to construct a method based on the minimization of the out-of-control expected delay for a given rate of false alarms. The problem we are dealing with is the time when our system is able to detect the γ-increase in the baseline rate.

Suppose we have X_i independent observations. With X we denote the number of 'normal' cases between two 'events'. Actually, X_i is the size of the set *i*. The events distribution can be approximated by a Poisson process with a parameter λ. The assumptions we make are the small event probability and the independence between the events. Thus, the observations can be supposed to follow an exponential distribution with a parameter $1/\lambda$. Under the baseline rate we have:

$$F_0(t) = 1 - e^{-\lambda_0 t} \quad (5.29)$$

And under the alternative hypothesis of the increased rate we have:

$$F_1(t) = 1 - e^{-\gamma\lambda_0 t} \quad (5.30)$$

With λ are denoted the events of interest. So with λ_0 we denote the average number of events of interest (i.e. the babies with congenital malformations) under the baseline rate. Our idea is to assign a score of 1 or -1 in each set X_i:

$$\text{Score}(X_i) = 1 \; if \; X_i < R$$

$$\text{Score}(X_i) = -1 \; otherwise$$

With R we denote the threshold value which in practice symbolizes the lowest 'normal' size of our set. A lower size than R, of a set, is an indication of an increased rate. Should this increase be statistically significant, it will be indicated by a sequence of consecutive sets. The value of R is predefined and in essence, is defined as *k-multiple of the expected value*:

$$R = k * E(x) \quad (5.31)$$

or under the baseline rate (null hypothesis) the (5.31) equation becomes:

$$R = k * c_0 \quad (5.32)$$

In the case of public health we are interested in the increase of the rate of a disease since a decrease is not our concern. Hence, we are dealing with the one-sided problem. For the one-sided case, the cumulative score's statistic is defined from the following:

$$S_0 = 0$$
$$S_i = \max\{S_{i-1} + score(X_i), 0\} \quad (5.33)$$

Our alarm rule is that an alarm is triggered as long as $S_i = n$.Where n is a fixed positive. The number of sequences from the beginning of our procedure until the alarm, is denoted by N (N = first i until the alarm). In essence, we can regard N as a random variable which represents the number of 'events' required to trigger an alarm. From this meaning we can easily jump to the conclusion that the expectation of N is the Average Run Length of our method.

Advising from the study of Munford(1980)[92] and proceeding with some algebra we conclude in the following expression of *E(N)* or *ARL*:

$$E(N) = \begin{cases} \dfrac{n(n+1)}{2p}, p = 0.5 \\ \dfrac{n}{2p-1} - \dfrac{1-p}{(2p-1)^2}\left\{1-\left(\dfrac{1-p}{p}\right)^n\right\}, p \neq 0.5 \end{cases} \quad (5.34)$$

Under Poisson process, ***p*** *which is the probability an observation is shorter than the decision threshold R*, is given by the following for the baseline and the increased rate of the tested hypothesis respectively:

$$H_0 : p_0 = 1 - e^{-k}$$
$$H_1 : p_1 = 1 - e^{-\gamma * k} \quad (5.35)$$

5.2.1 The Procedure

In order to optimize our procedure we aim at the out of control expected delay of the scheme described above. The assumptions we make is that the increase factor γ is detected and that the expected delay $E_0(N) = D_0$, is fixed. Then the procedure minimizes the out-of-control *ARL or* $E_1(N)$. Our method works following the steps described below:

1. We have an initial value of *n=1* and for this value we calculate $p_0 = \frac{1}{D_0}$.

2. For *n>1* we calculate the value of p_0 using the equation of G.Gallus et al.(1986)[50] which is defined as $P = [1 + D_0 - D_0 * P\ ]^{-1/n}$.We take a starting value of p_0 the value of p_0 for *n-1* and we act recursively for some iterations.

3. We calculate the value of *k* from $k = -\ln(1 - p_0)$.

4. We find the value of p_1 from *(5.35).*

5. We take the value of the expected delay $E_1(N)$ from equation *(5.34).*

6. We choose the appropriate parameters *k* and *n* for the minimum $E_1(N)$.

Note: It is concluded from simulations that the absolute minimum of $E_1(N)$ with respect to *n,* is the first minimum on *n.* So the procedure is stopped as soon as a value of $E_1(N)$ is found that is higher than the one obtained at a previous iteration.

Simulation

For the example in the sets method of the problem of congenital malformation in newborns, we had 400 births per month and one false alarm in 20 years of study. We are going to have our results for the 10 different baseline rates which corresponds to 10 different values of the λ_0 of the Poisson process. We are interested in detecting an increase in the normal rate of $p_1 = \gamma * p_0$ where $\gamma = 7$.

For this simulation we developed a code in S-plus and the Minitab statistical packages. This code is given in *Appendix B.*

From these data we have that we are studying 240 months and so the in-control expected number of malformations D_0 is:

Table 5.8: The expected number of malformations for different baseline rates

$\lambda_0 \times 10^{-4}$	D_0
1	10
2	19
3	29
4	38
5	48
6	58
7	67
8	77
9	86
10	96

The main idea is to choose the minimum out-of-control expected delay. In such a way we optimize our method. Thus, after our simulation and for the different values of the expected number of malformations we have the following table for the parameters of the optimal Cuscore method:

Table 5.9: The parameters of the Cuscore method for different baseline rates

$\lambda_0 \times 10^{-4}$	D_0	p_0	kappa	p_1	E_1	n
1	10	0,104167	0,110001	0,536990	1,86223	1
2	19	0,255630	0,295217	0,873374	2,45598	2
3	29	0,204474	0,228751	0,798358	2,82151	2
4	38	0,174905	0,192257	0,739668	3,17975	2
5	48	0,307976	0,368135	0,923994	3,43215	3
6	58	0,287570	0,339073	0,906847	3,54635	3
7	67	0,271499	0,316766	0,891104	3,65764	3
8	77	0,258387	0,298927	0,876621	3,76593	3
9	86	0,247404	0,284227	0,863248	3,87134	3
10	96	0,238017	0,271831	0,850852	3,97403	3

From this table we can derive the following graph from which we easily conclude that the delay, in detecting a statistically significant increase, is smaller for a rare (we are referred in the frequency of appearance) disease. Something expected. When a disease has a small rate of appearance in the examining population it is easy for the system to realize an increase, especially when this increase is large such as in this case *(γ=7)*.

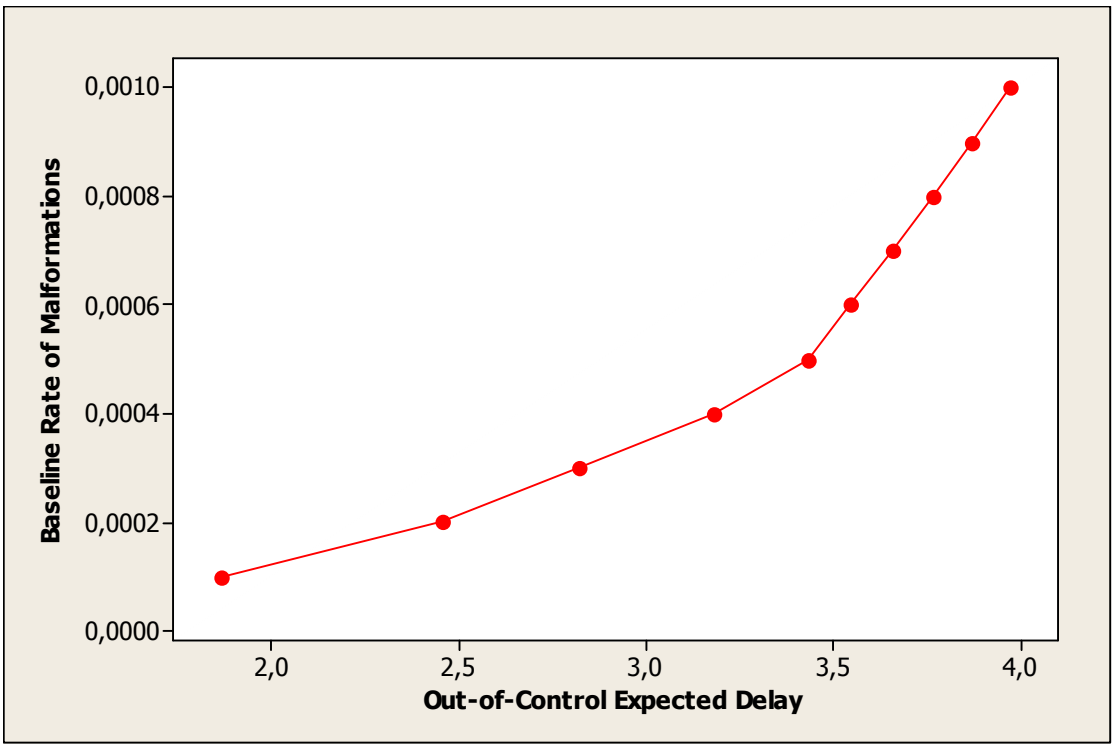


Figure 5.9: The expected delay until the alarm for the Cuscore method for different rates of newborns' malformations

5.3 Sets method vs Cuscore method

The sets method is appeared in the recent literature to be compared with other methods (e.g. the CUSUM scheme based on counts which we examine in the next chapter). Examples of this, is the study of Sego et al.(2008)[126] where two of the sets method's modifications are given (one of which is the Cuscore method). The evaluation of the performance of these methods was made with respect to the steady-state of the average run length instead of the ARL. The Cuscore test also is applied in the study of Chen and Froom(2003)[19] for the case of lymphoma and colon cancer death data.

In this section we compare the sets method and the Cuscore scheme based on the expected number of 'events' taken to detect the first true alarm after matching the rate of false alarms. In other words we are interested in the out-of-control expected delay in such an optimal way as the two methods were developed by the G.Gallus et al.(1986)[50] and G.Radaelli(1992)[102]. For the example above of the congenital malformations we calculate the out-of–control expected delay for the same parameters.

For the sets method, G.Gallus et al.(1986)[50] calculated the expected delay from the following equation:

$$E(N) = \frac{1 - p^n}{p^n (1 - p)} \quad (5.36)$$

For the example above we have the following table[5]:

Table 5.10: Comparing the Sets and the Cuscore method using the measure of the Expected Delay for the performance

					Expected Delay	
$\lambda_0 \times 10^{-4}$	p_0	kappa	p_1	n	*Cuscore Method*	*Sets Method*
1	0,104167	0,110001	0,536990	1	1,86223	1,86223
2	0,255630	0,295217	0,873374	2	2,45598	2,45598
3	0,204474	0,228751	0,798358	2	2,82151	2,82151
4	0,174905	0,192257	0,739668	2	3,17975	3,17975
5	0,307976	0,368135	0,923994	3	3,43215	3,52117
6	0,287570	0,339073	0,906847	3	3,54635	3,65962
7	0,271499	0,316766	0,891104	3	3,65764	3,79478
8	0,258387	0,298927	0,876621	3	3,76593	3,92649
9	0,247404	0,284227	0,863248	3	3,87134	4,05485
10	0,238017	0,271831	0,850852	3	3,97403	4,18005

[5] The code we used in S-plus for the Expected Delay of the Sets method is given in *Appendix C*

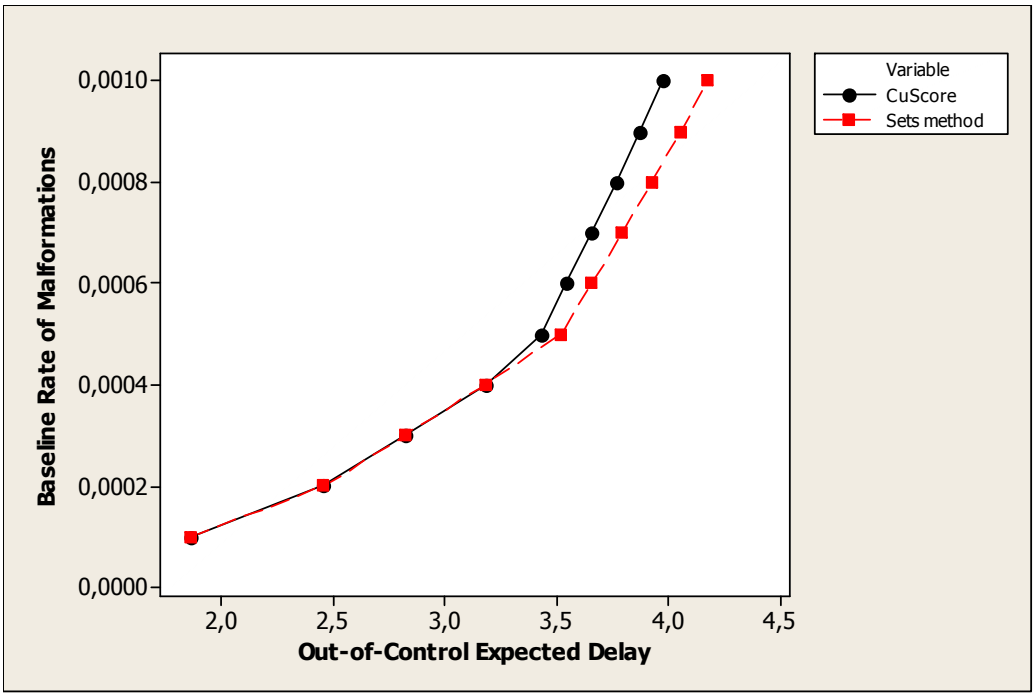


Figure 5.10: Comparing the performance of the Sets and Cuscore method using the measure of the expected delay until an alarm.

From the table and the graph above, we conclude that there is no difference between the Cuscore and the sets method for a small frequency of appearance of the malformation. There is no difference between the two methods for rare diseases. The difference is given for larger baseline rates of a disease.

Another interpretation of the above is that for a normal rate of 5×10^{-4} or higher the Cuscore method is more effective. The expected delay until our system triggers an alarm, in terms of the expected recorded events (congenital malformations in our example), is smaller for the Cuscore method than the Sets method. Especially for the case of the baseline rate 1 malformation per a thousand births, the difference is about **5%**.

As long as the normal rate is getting larger, we see that the gap between these two methods is bigger. Thus we should prefer the Cuscore method since it is much more effective, realizes the change in the 'normal' rate sooner than the sets method and it gives us reliable conclusions based on all the available data.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 6
Using the number of events in fixed intervals to study the Poisson process

In the previous section, in order to detect an increase in the baseline rate of a disease, we used the measure of the discrete time between two adverse events. In this section the time intervals are fixed and for these intervals we calculate the number of 'events'.

If the number of events is recorded for fixed time intervals, information on the process will be lost. As a result, the surveillance method will be suboptimal for detecting the change in the process as quickly as possible. Thus, there is only one reason to use fixed intervals. That is the practical restrictions of the reporting system.

Procedures for monitoring rare health events are based on sequential statistical methods for detecting a shift in the underlying disease rate. For fixed time intervals, a commonly used method is the Poisson CUSUM method. The Poisson CUSUM compares the recorded number of events in each time period with the expected number and uses the cumulated sum of deviations to form an alarm statistic.

A lot of comparisons of the Poisson CUSUM with other methods have been made. Such compared methods are the Sets method and the Cuscore of the previous chapter. As far as we are concerned about the recent literature of the period 2003-2012, the most commonly used is the CUSUM method for count data. The Poisson CUSUM performance is evaluated in Testik(2007)[153]. Especially, in the field of public health this method is widely spread and applied in different situations. For example, Limaye et al.(2008)[85] applied the counted CUSUM for the case of the hospital infections from children's hospital.

Additionally, several modifications of the Poisson CUSUM appear in the recent literature. Such an example is the Bernoulli CUSUM, in the Sego et al.

(2008)[126] and Sego (2006)[125] studies, which as we mentioned above is compared with the sets method and two of its modifications. It was found that the Bernoulli CUSUM is better, followed in order by the Cuscore and the sets method respectively. The same result for the most of the cases is derived by the study of Joner et al.(2008)[62], where the Bernoulli CUSUM is compared with a scan statistic for prospective surveillance.

To apply in theory the CUSUM scheme with count data is relatively easy, since we make the assumption that the mean number of events is steady for a whole time period. In practice though, this is not true. The mean number of events is likely to vary over time if we consider the factor of the population growth. That is the reason why a modification using weights is needed. That was shown in Shu et al.(2010)[131] and Shu et al.(2011)[132]. The weights correspond to the time the events occurred. More weights are assigned for the recent observations and fewer weights are assigned to the older observations. A comparison was made with the conventional CUSUM.

Another CUSUM method when the events to be monitored for an outbreak follow a Poisson distribution, is developed by Jonsson(2010)[64].

Further reading on comparisons of the Poisson CUSUM and other methods can be made through some interesting papers such as Barbujani and Calzolari(1984)[8], Pollak and Kenett(1983)[98], Gallus et al.(1986)[50], Chen(1987)[18] and Radaelli(1992)[102].

Most of these comparisons favor the Poisson CUSUM method but on the other hand there are proof for the opposite. For example in Chen(1987)[18] the sets method is evaluated with fixed intervals and it was found that it is a better method for a baseline rate equal or smaller than 5 cases per year. The evaluation was made with respect to the time delay measure until first alarm.

The conclusion mentioned above stands even for large increases in the baseline rate. The CUSUM scheme has shorter time delay, for a baseline rate larger than this value (5 per year). Further study of the same paper showed that the CUSUM scheme is preferred if it is applied at time intervals longer than one month.

A second scheme that has been developed is the method of the maximum value of the conditional likelihood ratios. This method is a complicated one and it demands a great amount of calculations and strong computing power. For

these reasons we are not going to present detailed information but only some general features in this study.

6.1 The Poisson CUSUM Method

The Poisson CUSUM scheme works through some parameters which are referred in the general features of the CUSUM method. Lucas(1985)[88] presented a general review of this method and gave us the appropriate tables to search for the parameters of our system according to the average run length measure.

For the CUSUM method, we usually use continuous data for our analysis. A first difference for the Poisson version is that we are dealing with discrete data. On the other hand, the interpretation of the results and the general conclusions (i.e. the planning of our method) is the same with the case of the continuous data.

To construct our method we use the Poisson distribution to model the number of counts observed per sampling interval. In order to use the Poisson CUSUM method in practical terms means that we have to make the assumption that we have the ability to record the number of counts in a fixed sampling interval. With the term ‘counts’ we mean the “events” of our sampling interval. In the public health field, as we mentioned above, we are interested in the case of the increase in the rate of a disease.

In this method we are dealing with the average run length as a measure of performance. With this term (average run length) we mean the average number of sampling intervals before an out-of-control signal. For the Poisson CUSUM the time between out-of-control signals is proportional to the *ARL*.

6.1.1 General Features of the CUSUM method

Cumulative sum methods are known for their ability to detect sudden changes in the mean of a variable. Especially when this change is quantitative large, the CUSUM scheme is very effective. For this method the assumption is

made that the variable exhibits no serial autocorrelation. The most common case is when the observations are normally distributed but in this case we are studying the Poisson distribution.

In the CUSUM method we are dealing with the parameters *h* and *k*. For our better understanding, though, we have to mention some information about *h* which represents our limit.

The *h* value is chosen based on a fixed value of rate of false alarms. High values of the limit *h* give us a low probability of a false alarm and a lower probability of detecting a real change in the studying mean (in the following we are going to see tables which define the values of *h* and *k* based on fixed values of ARL_0 for the Poisson version). It is common to use a value for *k* to be ½.If we assume that this is true then the ARL_0 is defined as follows:

$$ARL_0 = 2\left(e^c - c - 1\right) \quad (6.1)$$

where $c = h + 1.166$.

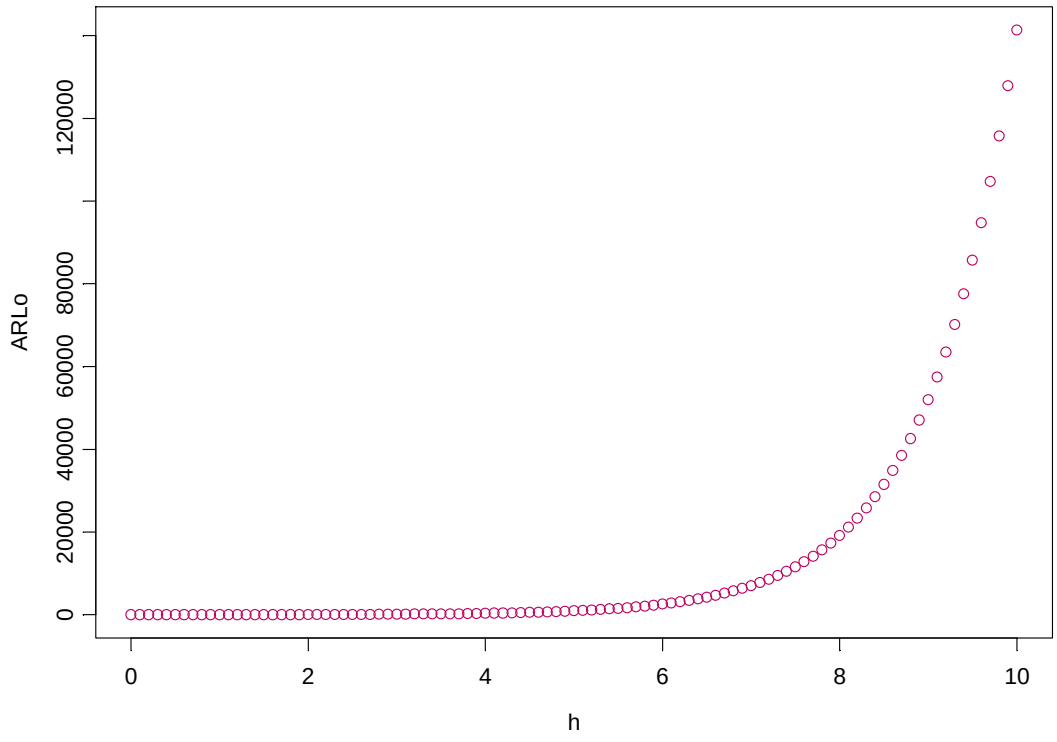


Figure 6.1: The performance of the Average Run Length until a false alarm according to the limit h of the CUSUM scheme

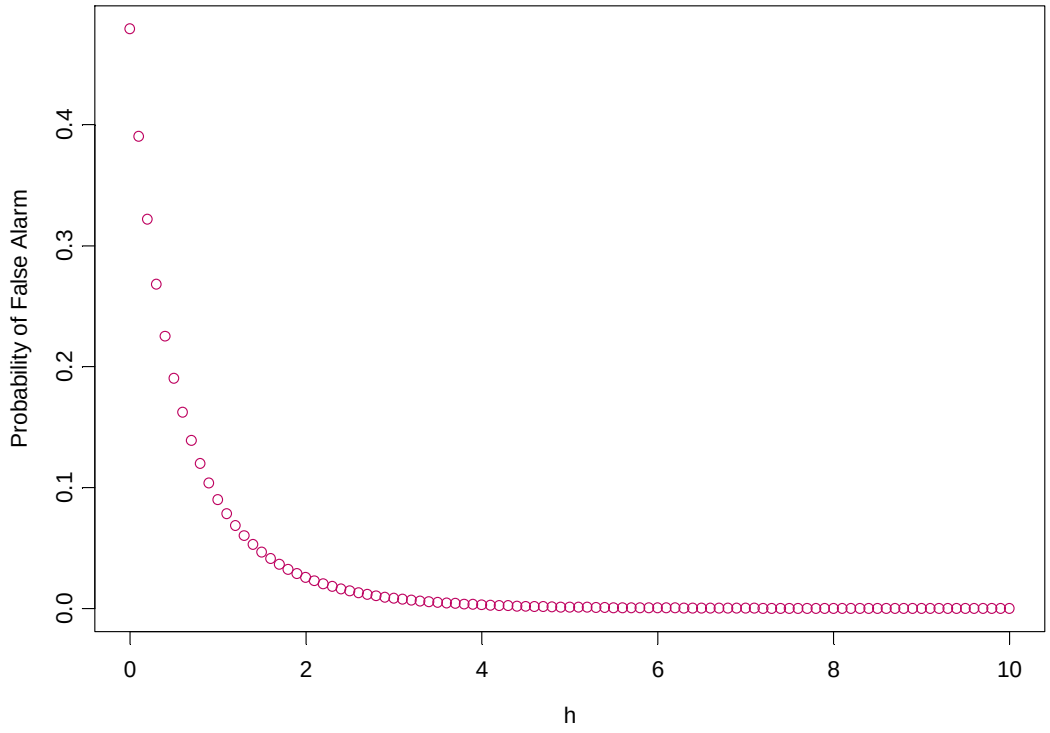


Figure 6.2: The performance of the PFA until an alarm according to the limit h of the CUSUM scheme

This is given by Siegmund(1985)[133]. As it is referred in Lawson and Kleinman(2005)[82] Siegmund's equation may be solved approximately for the threshold *h* as a function of the ARL_0, when *k=1/2*:

$$h \approx \left(\frac{ARL_0 + 4}{ARL_0 + 2}\right)\ln\left(\frac{ARL_0}{2} + 1\right) - 1.166 \quad (6.2)$$

When $k \neq 1/2$ this equation is formed as follows:

$$h \approx \left(\frac{2k^2 ARL_0 + 2}{2k^2 ARL_0 + 1}\right)\frac{1}{2k}\ln\left(2k^2 ARL_0 + 1\right) - 1.166 \quad (6.3)$$

6.1.2 Features of the Poisson CUSUM method

In this scheme we cumulate the difference between an observed value Y_i and a reference value *k*. If this sum equals or exceeds the decision interval value *h*, an alarm is triggered. For the Poisson CUSUM, the statistic is given below:

$$S_i = \max(0, Y_i - k + S_{i-1}) \quad (6.4)$$

Another feature of the Poisson CUSUM method is the Fast Initial Response(FIR) which was presented by Lucas(1985)[88]. Using this feature means that if a change takes place soon after our system's beginning, the CUSUM scheme is going to realize the change immediately. In practical terms it gives us a shorter average run length than the simple Poisson CUSUM method. For the case of the simple Poisson CUSUM the head start value for S_0 is set to be 0. On the other hand for the case of the Poisson CUSUM with the FIR feature, the head start value for S_0 is set to be equal to *h/2*. With such a head start our method will signal an alarm more quickly if our process is out of control soon after it started.

Another issue for this method is the proper choice of the values of *k* and the decision interval *h*. The value of *k* is chosen by the acceptable count rate and the count rate that is to be detected quickly. Using these two criteria for the choice of *k* value, leads us to build tables to choose the parameter *h*. The appropriate choice of these parameters, the tables and the exact procedure is given below via a conducted simulation.

6.1.2.1 The k-parameter

The *k*-value is chosen to be between the acceptable process mean (λ_0) and the mean level of counts that the CUSUM scheme is to detect quickly (λ_1). The means λ_0 and λ_1 are mean numbers of counts per sampling interval. The "acceptable" process mean' (λ_0) is the mean number of counts when the process

is in control. The "unacceptable" process mean (λ_1) is the mean number of counts when the process is out-of-control.

At this point we should mention that the desired value for λ_0 is zero. However, we usually do not use such a value for λ_0. Setting λ_0 with the value of zero, means that the CUSUM is designed with *h=1* and *k=0*.That means that for any occurrence of a count will give us an alarm. Therefore in practice λ_0, is chosen to be near to the current mean level. The reference value for the Poisson CUSUM could be selected to be close to:

$$k = \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)} \quad (6.5)$$

When $k \geq 1$, the *k* value will usually be rounded to the nearest integer.

6.1.2.2 The h-parameter

After determining the *k* value, the decision interval value *h* is chosen using proper tables. A proper choice of the value of *h* should give a large *ARL* when the counts are at the acceptable level and a small *ARL* value after the change has occurred.

These tables are given in the *Appendix D.1* and *D.2* (for the case with the FIR feature and the case without the FIR feature respectively) where we have the simulated table of the *ARL*'s, for the values of *h, k and* S_0 crossed with the mean of counts. In the case of the in-control state and the mean of normal counts, this represents the ARL_0 which we want to be a large number. In the case of the out-of-control state and the mean level of counts that the CUSUM scheme is to detect quickly, this represents the ARL_1, which we would like to be a small number.

Simulation

We suppose that in fixed time intervals we have a number of 'events' X that follows the Poisson distribution. Then we should have:

$X_i \sim Poisson(\lambda)$ for each of the $i=1,2,\ldots,n$ periods of interest.

Suppose now that the acceptable number of events is $\lambda_0 = 4$ and we have a sudden change in the mean of the events for about $\lambda_1 = 7$.The code we used in S-plus is given in the *Appendix D.3*. We generate from Poisson 20 values for each case which represent the number of events in the fixed time periods:

Table 6.1: The generated observations from Poisson distribution with a parameter 4 for the in-control state and 7 for the out-of-control state

Time Periods	$\lambda_0 = 4$	$\lambda_1 = 7$
1	2	6
2	3	10
3	2	8
4	2	6
5	2	6
6	5	4
7	5	10
8	3	10
9	4	7
10	1	14
11	1	2
12	5	9
13	2	12
14	8	15
15	4	9
16	2	6
17	3	4
18	5	5
19	7	6
20	6	2

❖ *With the FIR feature*

For these mean numbers of the two states of control, we have that the *k*-value should be:

$$k = \frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)} = \frac{3}{1{,}95 - 1{,}39} = 5{,}35$$

Hence, the value of *kappa is 5*. Moreover we are interested in a value of *h* equal or larger than seven. So we have the ARL's for the 2 cases and we conclude that a proper value could be the one for *h=10*.

The average run length until a false alarm is 397 and the average run length until our system realizes the change is about 3 or 4 periods.

Table 6.2: The table with the Average Run Lengths for the in-control and out-of-control state of the Poisson CUSUM scheme (using the FIR feature)

			Mean as a multiple of *k*	
h	*k*	$S_0 = h/2$	*4/5=**0.8***	*7/5=**1.4***
7	5	4	94.9	2.37
10	5	5	397	3.35
15	5	8	3630	4.36

With the values of *k=5* and *h=10* we have the following graph including the FIR feature where we have an alarm with a delay of 3 time periods after the change:

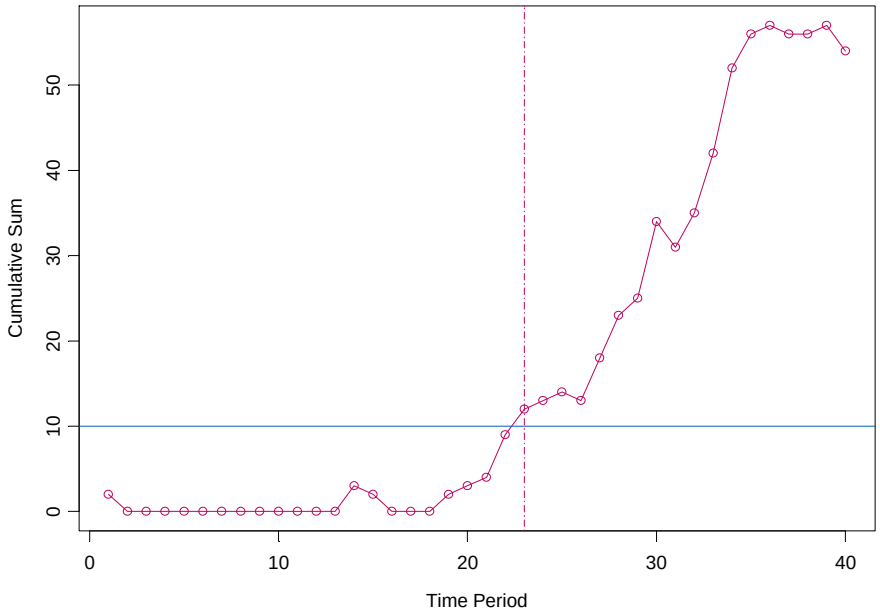


Figure 6.3: The Poisson CUSUM scheme using the FIR feature and limit h=10. Alarm at the 23rd observation

❖ *Without the FIR feature*

In this case where the FIR feature is not included and for the value of the parameter of *kappa* which we found above, we have the following table:

Table 6.3: The table with the Average Run Lengths for the in-control and out-of-control state of the Poisson CUSUM scheme (without using the FIR feature)

			Mean as a multiple of *k*	
h	*k*	$S_0 = 0$	4/5=**0.8**	7/5=**1.4**
7	5	0	108	4.09
10	5	0	422	5.59
15	5	0	3740	8.09

Therefore we choose a value of *h* to be 10.For this value as before we have the following conclusions.

The average run length until a false alarm is 422 and the average run length until our system realizes the change is about 5 or 6 periods which is a larger number of periods compared with the previous case of the FIR feature.

We see that the procedure with the FIR feature gives our Poisson CUSUM an advantage. That is the quicker realizing of the change in the mean and in terms of the time delay it is more effective. The Poisson CUSUM is given below:

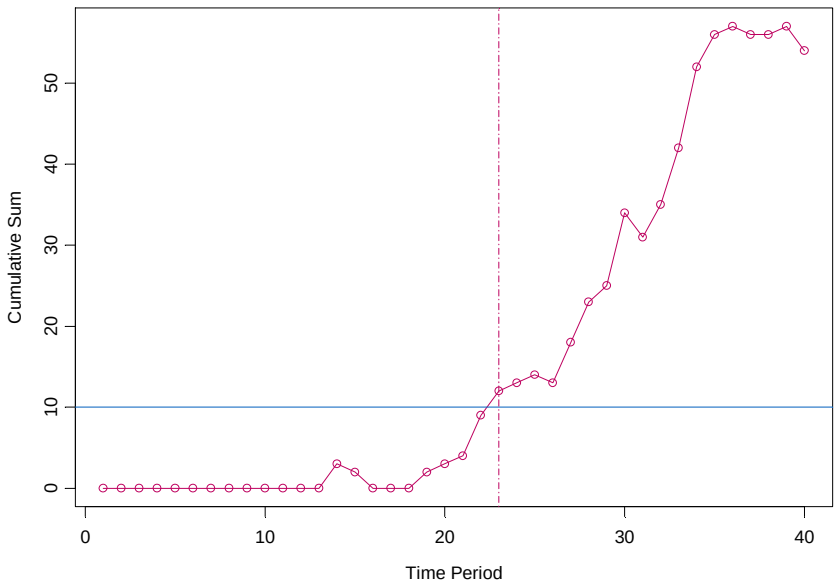


Figure 6.4: The Poisson CUSUM scheme without using the FIR feature and limit h=10. Alarm at the 23rd observation

6.1.2.3 Poisson Approximation to a Normal Process

If the in control value of λ is larger or equal than 2, we can transform the counts to a standard normally distributed random variable using the transformation of Rossi et al.(1999)[120]:

$$x = \frac{p - 3\lambda + 2\sqrt{\lambda p}}{2\sqrt{\lambda}} \quad (6.6)$$

Where p is the observed count, λ is the expected count and x is our normal random variable. For values smaller than $\lambda<2$ our results and conclusions cannot be reliable as Rogerson and Yamada(2004a)[117] showed.

Another issue for this approach is that the delay is shorter with the Poisson CUSUM. After the transformation of the data, the delay of the detection of an outbreak becomes longer in the CUSUM scheme. Thus this transformation leads

to longer time delays of detection of a change. That was shown in Hawkins and Olwell(1998)[55].

6.2 The maximum value of the Conditional Likelihood ratios

Another approach for the case of the fixed intervals is the maximum value of the conditional likelihood ratios as the alarm statistic (proposed in Lie et al.(1993)[84]). In this case we are based on a series of sequential tests and interested in the characteristics of the whole process.

Lie et al.(1993)[84] presented a sequential binomial likelihood ratio test of the probability that an infant has Down's syndrome. The procedure was called the *γ-method* and was based on a parametric model for maternal age specific proportions of Down's syndrome cases, assuming that a certain fraction of the cases is attributable to causes unrelated to maternal age.

However what is interesting in Lie's paper is the comparison with the Poisson CUSUM method. According to that paper, the Poisson CUSUM has been shown to be 44% slower in detecting a moderate increase in Down's syndrome risks that occurs additively over all maternal ages. In most cases the CUSUM method seems to be 20 to 30 per cent less sensitive to additive increases in terms of the average run length.

Chapter 7
The Historical Limits Method

The detection of patterns in the occurrence of diseases and other health events presents an important challenge to public health surveillance. For that purpose a window based method was firstly proposed by Stroup et al. (1989)[47] & (1993)[46].

Aberrations in usual distributions of disease incidence may provide an early signal of an epidemic or may provide clues to important risk factors associated with the occurrence of a disease in a particular space and time. Detection of an out-of-control state of a disease is the subject of our study in this case and to achieve that goal we are based on an existing data set.

The particular study is based on the weekly reports of up to 50 diseases made by the state health departments and which are submitted by the National Notifiable Diseases Surveillance System (NNDS) of the Centers for Disease Control (CDC).These reports are disseminated in the Morbidity and Mortality Weekly Report (MMWR) and are available to epidemiologists, clinicians and other public health professionals.

There are some studies which are dealing with the prospective case of this method. A very interesting study is the one of Rigau-Perez's. Rigau-Perez et al.(1999)[107] used as historical data, the reports from 5 years in which it was not reported a dengue outbreak in the geographic area of Puerto Rico. From each of these years it was chosen a three month window (15 weeks total) with the fewest dengue cases in Puerto Rico. By comparing all the possible three month windows it was decided the use of the window from April to June as this "window" had the less 'events' in the particular area of study. In this study the mean was the measure of comparison and so it was calculated from the total of 60 weekly reports. The results of this retrospective application were used to develop a surveillance method with predictive capabilities. A comparison of these results was made with the data gathered after the period of interest

(current period) and tests were made in order to find out if there is a significant difference between the results of this comparison and the outcome of the real data. In the end these comparisons conclude in a single surveillance method (with high sensitivity and specificity) which functions as an indicator for evaluation of dengue prevention and controlling a probable outbreak.

Another study for the prospective case should be the one of Wharton et al.(1993)[162] who used data from the NNDSS for a 4 month period.

In Stroup et al. (1993)[146] three algorithms were used for estimating the standard error of a simulated ratio (from a known distribution). Those are the bootstrap, the jackknife and the delta methods. For the evaluation of these methods for the best given variance, a model is used by which results can be compared with true or model-simulated values. Moreover, the three methods were applied to real data from the NNDS and measured performance by the epidemiologic confirmation of increased activity.

Conclusions showed that the simple bootstrap which is based on random sampling from the fifteen past data, and the jackknife procedures are not the most appropriate. They produce exaggerated low estimates of the variance used for the ratio in our *graph* method (which in fact means that we have a little difference between the past and the current value). These confident estimations have as a result values for the ratio very close to 1 and this underestimation will lead to a very sensitive method with often out-of-control states. On the other hand, the delta method produced the best estimate of the true limits in our graph method.

Based on a moving window Shore and Quade (1989)[130] proposed the short memory method and compared it with the Poisson CUSUM method. Another example of the use of the window based method is the detection of increased γ-radiation levels in Sweden. In this case two consecutive 24-hours periods are compared by the Swedish Radiation Protection Institute (Kjelle (1987)[68]).

Further research has been conducted the last few years on the Historical Limits Method. In fact this method has been the object of comparison with other methods. An example is the paper of Choi et al.(2010)[24] where the historical limits method is compared with six other methods (one of them is the Serfling method which is described in the chapter 9.3 of this study).Measures of

evaluation for this comparison is the sensitivity, the specificity, the positive predictive value and the short time lag. These methods were applied in simulated and real data for an overall view. The poor performance of the HL method is mentioned compared to other methods.

The seasonality of a disease is the main interest of the study of Pelecanos et al.(2010)[96]. The seasonal factor in many cases is related with a possible outbreak but constructing a system based on a seasonal outbreak will lead to inconclusive and false results in the future. Different algorithms for seasonal data are compared in this study.

7.1 The Time Window

These reports are referred to national data and our goal is to use as short a time period as possible for weekly publication in order to empower the usefulness of our surveillance method. However, there is an important problem. That is the variability in the weekly reports caused by factors irrelevant with our process such as the time delay of the reporting because of outbreaks. For avoiding the contingency of the instability of our results, we choose a 4-week time window.

7.2 The method

The purpose of our method is to facilitate the analysis of surveillance data and to offer in other sources of information. The method may not be useful for conditions with long-term historical trends and it is more effective if the base-line rate of a disease is completely unknown.

The reports, mentioned above, consist of an enormous amount of data which need a clear and effective statistical analysis. For this reason, a bar graph with the ‘ability’ to detect important changes in the baseline rate was developed. This graphic method was developed to describe the comparison between the current and the past data in a simple optical way. The second reason for developing such a graph is to highlight and draw the attention of the reader in the indication of changes in long-term trends or epidemics.

Reporting cases to an organization who observes the control of diseases (e.g. CDC) is a simple process but there are some problems. For example reporting cases may vary by factors unrelated to the disease process (e.g. reporting practices or time of month).

In order to reduce such variability from these causes, we aggregate disease reports over a month period. The choice of the 4-weeks' "current period" is not random. It is extracted from the fact that weekly fluctuation in disease reporting is noticed (it is usually due to irregular reporting rather than to disease incidence).

Let now x_0 be the number of cases of a given disease reported to an organization in the four week period ending with the current week. We also have the observations $x_1,...,x_{15}$ which represent 15 previous totals and from which we take a baseline report. These observations represent the values of a three month window (the corresponding month and the surrounding months) in each of the five previous years.

In such a way we create a baseline rate. Each time we compare the number of cases we are interested in with these 15 previous observations (or the baseline rate).

In the following table we have a visual view of the time window of interest where

x_0 : is the number of cases of the current period (February 2012).

$x_1,...,x_{15}$: is the historical observations of a three month window from which we take our confidence interval.

Table 7.1: The table with the observations in the time-window of interest

	January	**February**	**March**
2012		x_0	
2011	x_1	x_2	x_3
2010	x_4	x_5	x_6
2009	x_7	x_8	x_9
2008	x_{10}	x_{11}	x_{12}
2007	x_{13}	x_{14}	x_{15}

Hence, the question: “is the number of cases of this month different from last month” is being transformed to the question: “is the number of cases this year different from last year”. With this transformation of our question we avoid the variance caused by seasonality. Seasonal diseases are observed to have an outbreak in a particular season. These seasonal increases of a rate of a disease should not be taken into account because an epidemic is not based on a temporary increase in the normal rate but on a long-term and at least a more permanent increase.

An assumption made in our method is that $x_1,...,x_{15}$ and x_0 are independent random variables with the same distribution function. For most diseases, the three month window produces data which satisfy this assumption. Thus, we calculate a two sided confidence interval for the expected number of cases for a four week period. We compare this confidence interval and the observed current value x_0 and we conclude whether the disease process is out of control for the current month or not.

From the epidemiological and statistical view of this matter, we could say that we are interested in what sense the expected number of ‘events’ is approaching the observed (real) value of x_0 and if this number of cases is exceeding the control limits. These questions are of significant importance since we do not know if a value from the previous years has exceeded the in-control state. To solve this problem we use the law of large numbers and so we expect that in large samples the estimation of a mean (or even a median) calculated from the sample, is very likely to approach the real value of the parameter for the entire population. The problem of the window based method is that taking a sample from the previous 15 months is not a large sample. To continue with our method we use a ratio of *y*:

$$y = \frac{x_0}{m},$$

where *m* is a measure of central tendency of the 15 baseline values. For this measure we evaluate the mean or the median of the 15 baseline values, depending in our desired level of sensitivity of the method with regard to aberrations in the baseline.

Summarizing up, this is a parametric method which uses the 15 baseline values to compute a normal theory confidence interval, using the mean (or the median) of the baseline period in the denominator.

7.3 Diseases for which our method is appropriate

Our method is most appropriate for diseases that do not appear a significant level of trend in the past (historical data). Additionally these diseases should occur often enough so that a few cases will not be the reason of having a signal of alarm in our surveillance. If the data are not preanalyzed for trend and period effects and the variance of the current cases is assumed to be the same as the variance of the observations of historical data, our *graph* may be less powerful.

7.4 The Graph

In order to develop our graph, we have to think of the arithmetic ratio of current to historical incidence. In the axis of *x* we have the value of the ratio and on the axis of *y* we have the diseases of interest. We also have a vertical axis on the value of *1* of the ratio axis. This value actually means that there is no change in the events of the current period compared with the last 5 year's data.

In the case of spatial surveillance, which we are going to examine in the third part of our study, the *y-axis* does not represent the diseases but different geographic areas (e.g. countries, states, cities etc) since our graph presents the evolution of a particular rare disease in different countries or cities. In fact, we are interested in the detection of an outbreak in an area or in the spatial clustering of a particular disease. More details for spatial surveillance of public health are shown in the 3rd part of this study.

The point where the hatched area begins is based on the mean and two standard deviations of the 15 values of events occurred in the past. These are the

two factors from which we develop a statistical formation of a confidence interval.

We give the upper and lower control limits for a ratio of a disease. Additionally, two figures are given for each case. These are the deviation bar charts of rare reported notifiable diseases (<1000 cases reported during the preceding year), from the Centers for Disease Control and Prevention's Morbidity and Mortality Weekly Report for the United States of America:

- For the case of the significant increase, we have an alarm if the number of 'events' of the current period is exceeding the limit based on the past 'events'. This limit is defined as $m+2s$.Thus, our alarm rule should be stated as *we have an alarm if:*

$$x_0 > m+2s \Leftrightarrow$$
$$\frac{x_0}{m} > 1+\frac{2s}{m} \Leftrightarrow$$
$$y > 1+\frac{2s}{m}$$

An alarm in the case of an increase in the rate notifies us about a foreseen outbreak of a disease and that is the case where the most epidemiologists are of interest.

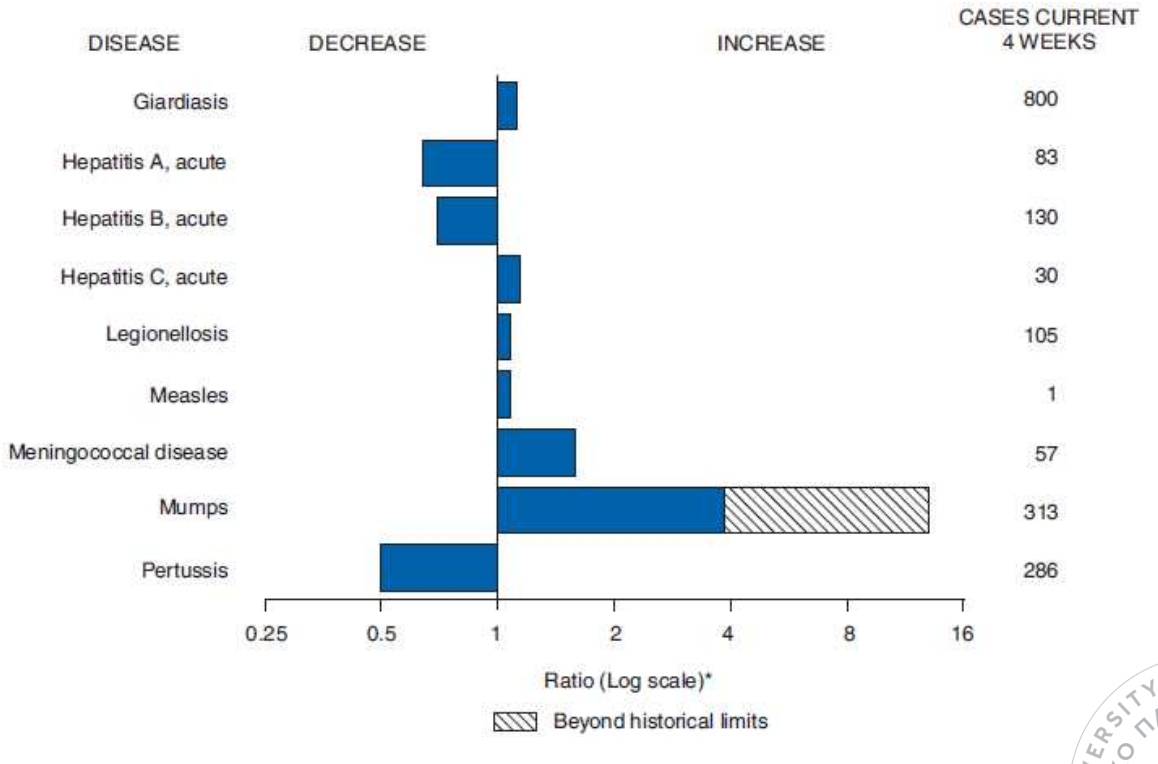


Figure 7.1: Selected notifiable disease reports, comparison of provisional 4-week totals December 26, 2009, with historical data. Ratio of current 4-week total to mean of 15 4-week totals (from previous, comparable and subsequent 4-week periods for the past 5 years). The point where the hatched area begins is based on the mean and two standard deviations of these 4-week totals.

❖ For the case of the significant decrease, we have an alarm if the number of 'events' of the current period is lower than the limit based on the past 'events'. This limit is defined as $m-2s$. Thus, our alarm rule should be stated as *we have an alarm if:*

$$x_0 < m-2s \Leftrightarrow$$
$$\frac{x_0}{m} < 1-\frac{2s}{m} \Leftrightarrow$$
$$y < 1-\frac{2s}{m}$$

An alarm in the case of a decrease in the rate notifies us about the effectiveness of some measures taken in order to tackle with an outbreak of a disease in the past. Its role is mostly retrospective and gives us a clear view if we have faced an outbreak or not.

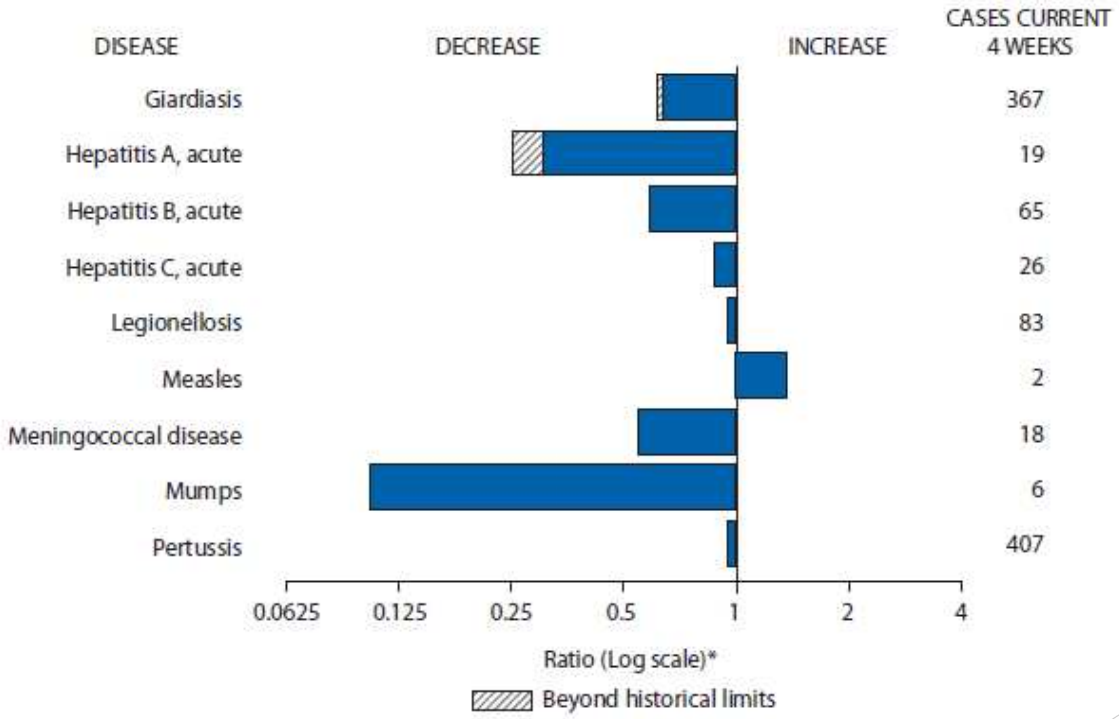


Figure 7.2: Selected notifiable disease reports, comparison of provisional 4-week totals January 28, 2012, with historical data. Ratio of current 4-week total to mean of 15 4-week totals (from previous, comparable and subsequent 4-week periods for the past 5 years). The point where the hatched area begins is based on the mean and two standard deviations of these 4-week totals.

The parameters *m* and *s* are the mean and the standard deviation of the baseline data. Then our confidence interval or the upper and lower limits are best described by the following expression for the ratio *y* of our interest:

$$\left[1-\frac{2s}{m},1+\frac{2s}{m}\right]$$

Note: *This method does not adjust for non-normal and serially correlated data. So using this method means the acceptance of the assumptions of the normally distributed data and the uncorrelated number of cases between periods.*

7.5 Problems

The main disadvantage of the window-based methods is that they are suboptimal procedures. For example the ability to detect a gradual change is low (Svereus (1995)[49]), if we proceed in the comparison of two consecutive moving windows of fixed lengths. Additionally, using time windows will have as a result a great loss of information as only the data referred in our period of interest will be taken into account.

Surveillance data are reported sequentially in time. That is the reason why they may not satisfy the assumptions necessary for usual time series analyses. These problems appear especially for incidence data for which the numbers of reported cases are subject to seasonal effects and reporting delays. If at least one of our assumptions is not accepted, then the method used to set the limits of our graph may be affected. In fact, if there is limited knowledge for the empirical performance of the method in the absence of these two assumptions, this affection is greater.

On the investigations of these problems there are a lot of useful studies including Kafadar and Stroup(1992)[65], Efron and Tibshirani(1986)[32], Bose A.(1988)[9], Liu(1988)[86], Kunsch(1989)[75].

What we should have in mind, is that there is no single method which can be used to detect all epidemics or all types of aberrations. Incomplete or inaccurate reporting is expected to affect our method but even in this case these problems can be turned out to be useful considering the fact that we can use the available data to detect trends or patterns. All the data, followed by the proper evaluation, are useful and that is a rule that in the field of public health should be unbreakable.

Chapter 8
Optimal Surveillance Methods

In the previous chapters, we examined some suboptimal surveillance methods. In this chapter we are dealing with methods which can satisfy at least one optimality criterion. Such an optimal criterion as we saw in *chapter 3* is the *minimization of the expected delay for a fixed probability of a false alarm.* We examine this criterion mostly for two reasons. It applies in several studies and in the field of public health is a priority.

In this chapter we are dealing with likelihood methods which can fulfill properties of the system expressed by the optimal criterion mentioned above. Such methods are the Likelihood Ratio and the Shiryaev-Roberts method.

8.1 Likelihood Ratio Method

Frisén and de Maré (1991)[46] showed that the minimization of the expected delay for a fixed probability of a false alarm leads to the full Likelihood Ratio method. This method is fully described from the following. The partial likelihoods are denoted by *L(s,t)* and are defined as follows:

$$L(s,t) = \frac{f(x_s \mid C(s))}{f(x_s \mid D(s))}$$

The *full likelihood* is a weighted sum of the partial likelihoods as we are going to see later.

The Likelihood Ratio method has an alarm set consisting of those X_s for which the full Likelihood Ratio exceeds a limit. Then, the time of an alarm for the Likelihood Ratio method can be expressed as the first time that the posterior probability of a change exceeds a limit ***K***:

$$t_A = \min\{s; P(\tau \le \mathrm{s} \mid \mathrm{X}_s = x_s) > K\}$$

The first time that the full Likelihood Ratio exceeds a time-varying alarm limit is an equivalent way of describing the time of an alarm for the likelihood ratio method:

$$t_A = \min\left\{ s; \sum_{t=1}^{s} w(s,t) * L(s,t) > \frac{\Pr(\tau > \mathrm{s})}{\Pr(\tau \le \mathrm{s})} \frac{K}{1-K} \right\}$$

where $w(s,t) = \dfrac{\Pr(\tau = \mathrm{t})}{\Pr(\tau \le \mathrm{s})}$ .The $w(s,t)$ are the weights which correspond to the partial likelihoods $L(s,t)$ and K is the limit at the decision time s. Therefore, for the time of an alarm we have the following rule:

$$t_A = \min\left\{ s; \sum_{t=1}^{s} \frac{\Pr(\tau = \mathrm{t})}{\Pr(\tau \le \mathrm{s})} * L(s,t) > \frac{\Pr(\tau > \mathrm{s})}{\Pr(\tau \le \mathrm{s})} \frac{K}{1-K} \right\}$$

Thus, in order to solve this problem we have to make only one important assumption. That is the distribution of τ. We assume that τ is a random variable which follows a particular distribution. For this variable it is common to use a *geometric* distribution.

A lot of studies have as a subject the Likelihood Ratio method for a change in several distributions. One example is the Poisson approach of the Likelihood Ratios as it is going to be described below. Another example is the case of a normal distribution. In this case, the *LR method* is optimized for the values of the change size μ and the change intensity v is used in the alarm statistic. For the case of normal distribution we have an alarm at:

$$t_A = \min\left\{ s; \sum_{t=1}^{s} \Pr(\tau{=}\mathrm{t}) * \exp\left\{ \frac{t\mu^2}{2} \right\} \exp\left\{ \mu \sum_{i=t}^{s} X(i) \right\} > \exp\left\{ \frac{(s+1)\mu^2}{2} \right\} \Pr(\tau > \mathrm{s}) \frac{K}{1-K} \right\}$$

8.2 Shiryaev-Roberts method

In the method described above (Likelihood Ratio method) the conditional likelihood ratios are weighted according to the distribution of the change point.

In this case we examine the "Shiryaev–Roberts" method where the partial likelihood ratios have equal weights. That is the main difference between these two methods. The *SR method* is the limit of the *LR method*:

$$\sum_{t=1}^{s} w(s,t)L(s,t) > K$$

when *v* tends to zero, since both the weights $w_s(t)$ and the limit K tend to constants. Shiryaev and Roberts suggested that an alarm is triggered at the first time *s* for which:

$$\sum_{t=1}^{s} L(s,t) > K$$

where K is a constant. Thus, the time of an alarm, enhanced with the optimality criterion, for the Shiryaev-Roberts method can be expressed as follows:

$$t_A = \min\left\{s; \sum_{t=1}^{s} L(s,t) > K\right\}$$

In practical terms the Shiryaev-Roberts method can be used as an approximation of the Likelihood Ratio method. If we consider a low intensity of a shift, the parameter of the geometric distribution is close to 0 and as a result the weights (of the LR method) are becoming (almost) constants. So there is no need in using them in our alarm statistic and therefore we achieve this approximation. Many studies have examined this matter. For example Frisén and Wessman (1999)[48] demonstrated that the SR method is a good approximation for intensities *v* less than **0.20** in the case of a change in the mean of a normal distribution. In the same study another property is used. That is the *constant Predictive Value*.

8.3 Linear approximations of the Likelihood Ratio method

A lot of studies are dealing with the matter of linear approximations of Likelihood Ratio method. We study different kind of linear approximations of the LR method mainly for two reasons. Firstly, we need to construct a method which is easier to use and analyze, but has similar good properties as the LR method. Another reason is to get a tool for the analysis of approximate optimality of other methods. Different approximations might be of interest for different situations.

An approximation, which is denoted by ***LinLR*** is achieved by a Taylor approximation of the alarm function and gives an alarm for the first *s* for which:

$$\sum_{t=1}^{s} W_{LinLR}(s,t) * x(t) > L_{LinLR}$$

Other approaches could be the *Exponentially Weighted Likelihood Ratio (EWLR)* method and the *Exponentially Weighted in Likelihood Ratio (EWLinLR)*. Further explanations are given in the study of Frisén (2003a)[37].In the appendix of this study are given all the necessary calculations which lead in the approximations mentioned above. Also some interesting simulations are given in Frisén and Sonesson (2002)[47].

8.4 The *LR* and *SR* methods for a Poisson process

The Likelihood Ratio method and the Shiryaev–Roberts method may be applied for the case of a positive shift in a Poisson process too as it is shown in the study of Sonesson and Bock(2003)[140]. Particularly, the construction of these two methods can be applied both in the case when *data are represented by the time between events* and when *data are represented by the number of events in fixed intervals* (as they are described in the chapters 5&6).

8.4.1 Data are represented by the time between events

Here we shall derive the Likelihood Ratio method and the Shiryaev–Roberts method for the case of an increased rate of incidence when the cases are assumed to follow a Poisson process.

If the cases are distributed by Poisson with a parameter *v*, the time intervals *X* follow an exponential distribution with a parameter *1/v*. Now imagine that we have to deal with a shift in the intensity of our process from v_0 to v_1, considering our two states $D(s) = \{\tau > s\}$ and $C(s) = \{\tau \le s\}$ respectively.Remember that for this case we make the assumption that the timescale for τ follows the number of events. So, for exponentially distributed time intervals denoted by *X*, we should have the following probability density functions:

$$f(x_i;1/v_o) = \frac{1}{1/v_o}\exp\left\{-\frac{x_i}{1/v_o}\right\}$$

And

$$f(x_i;1/v_1) = \frac{1}{1/v_1}\exp\left\{-\frac{x_i}{1/v_1}\right\}$$

Thus our *likelihood functions* should be for each case:

$$f(x;1/v_o \mid D) = v_o^{\,s-t+1}\exp\left\{-v_o\sum_{i=t}^{s}x_i\right\}$$

And

$$f(x;1/v_1 \mid C) = v_1^{\,s-t+1}\exp\left\{-v_1\sum_{i=t}^{s}x_i\right\}$$

Thus *the partial likelihood* should be:

$$L(s,t) = \left(\frac{v_1}{v_o}\right)^{s-t+1}\exp\left\{-v_1\sum_{i=t}^{s}x_i + v_o\sum_{i=t}^{s}x_i\right\}$$

8.4.1.1 Time of an Alarm for the LR method

The time of an alarm for the Likelihood Ratio method and a constant *K* is,:

$$t_A = \min\left\{ s;\ \sum_{t=1}^{s} \frac{\Pr(\tau=t)}{\Pr(\tau\le s)} * L(s,t) > \frac{\Pr(\tau>s)}{\Pr(\tau\le s)} \frac{K}{1-K} \right\}$$

$$\Leftrightarrow$$

$$t_A = \min\left\{ s;\ \sum_{t=1}^{s} \Pr(\tau=t) * \left(\frac{\nu_1}{\nu_o}\right)^{s-t+1} \exp\left\{ (-\nu_1+\nu_o) \sum_{i=t}^{s} x_i \right\} > \Pr(\tau>s) \frac{K}{1-K} \right\}$$

8.4.1.2 Time of an Alarm for the SR method

For the same case of an increased rate of incidence when the cases are assumed to follow a Poisson process, *the time of an alarm for the Shiryaev-Roberts method* and a constant *C* is:

$$t_A = \min\left\{ s;\ \sum_{t=1}^{s} \left(\frac{\nu_1}{\nu_o}\right)^{s-t+1} \exp\left\{ (-\nu_1+\nu_o) \sum_{i=t}^{s} x_i \right\} > C \right\}$$

8.4.2 Data are represented by the number of events in fixed intervals

For the second case we derive the Likelihood Ratio method and the Shiryaev–Roberts method when our observations consist of a number of events *X*, recorded in fixed intervals of length *k*. The number of events in fixed intervals of length *k* follows the Poisson distribution. So the density function is:

$$P(x_i;\nu_o) = \exp\{-\nu_o\} \frac{\nu_o^{x_i}}{x_i!}$$

Thus our *likelihood function for the case of the in-control state* should be:

$$P(x;v_o \mid D) = \exp\{-k(s-t+1)v_o\}\frac{v_o^{\sum_{i=t}^{s} x_i}}{\prod_{i=t}^{s} x_i!}$$

With the same way we derive *the likelihood function for the case of the out-of-control state*:

$$P(x;v_1 \mid C) = \exp\{-k(s-t+1)v_1\}\frac{v_1^{\sum_{i=t}^{s} x_i}}{\prod_{i=t}^{s} x_i!}$$

Thus *the partial likelihood* should be:

$$L(s,t) = \left(\frac{v_1}{v_o}\right)^{\sum_{i=t}^{s} x_i} \exp\{k(s-t+1)(-v_1+v_o)\}$$

8.4.2.1 Time of an Alarm for the LR method

For that case, *the time of an alarm for the Likelihood Ratio method* and a constant *K* is:

$$t_A = \min\left\{s;\ \sum_{t=1}^{s} \Pr(\tau=t)*\left(\frac{v_1}{v_o}\right)^{\sum_{i=t}^{s} x_i} \exp\{k(s-t+1)(-v_1+v_o)\} > \Pr(\tau>s)\frac{K}{1-K}\right\}$$

8.4.2.2 Time of an Alarm for the SR method

For the same case of the number of events in fixed intervals of length *k*, *the time of an alarm for the Shiryaev-Roberts method* and a constant *C* is:

$$t_A = \min\left\{ s;\ \sum_{t=1}^{s} \left(\frac{v_1}{v_o} \right)^{\sum_{i=t}^{s} x_i} \exp\{k(s-t+1)(-v_1+v_o)\} > C \right\}$$

8.5 Conclusions

In both cases the Likelihood Ratio and the Shiryaev–Roberts method are preferable to the suggested methods for the cases of chapters 5 & 6. The *LR* and the *SR* methods fulfill optimality criteria when the methods of the previous chapters are suboptimal methods. The problems surely are handled better using the ratio of the likelihood functions, as the expected delay will be shorter for a fixed value of the probability of a false alarm.

Table 8.1: The methods, their alarm functions, the number of their parameters and optimality

Method	Alarm function of L(s,t)	Number of parameters in the Alarm functions	Optimality
Likelihood Ratio	$\sum_{t=1}^{s} w_s(t) L(s,t)$	*2*	$\min\{E(t_A - \tau \mid t_A > \tau)\}$ *for a fixed* $P(t_A < \tau)$
Shiryaev-Roberts	$\sum_{t=1}^{s} L(s,t)$	*1*	*Same as in LR method* for $v \to 0$

Sometimes the Likelihood Ratio method (which means the use of the posterior distribution) is named “the Bayes method” while the Shiryaev-Roberts method is considered a frequentistic one. That is not fully true. In fact no especially Bayesian assumptions are necessary for the LR method. The identification of such a method as a Bayesian one depends on the situation whether the distribution of τ is considered as a “prior” probability, an observed frequency distribution or just reflects which situation optimality is desired.

In recent literature there are several applications of the LR method in public health because of its optimal properties. In the study of Andersson(2003)[3] a comparison is given between the LR method and the non-parametric approximation of the LR method for the case of detecting influenza epidemics. Another comparison is also available in Andersson(2004)[4] between the LR method and the maximum likelihood ratio method.

For the case of Chang(2008)[13] the SR method applied in monitoring surgical performance is compared with two different CUSUM schemes.

The optimality of methods is a very interesting subject for further research. Expressing methods through likelihood functions gives us a whole new field of study in order to link methods with optimality criteria. An interesting study in this area is the one of Frisén(2007)[39] & (2009)[40].

PART II

Detection of Increased Rates of Incidence in a Non Poisson Process


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 9
Processes with Time Dependencies

In the previous chapters we are dealing with methods which can be applied in a Poisson process. Actually that is our main assumption. There are, though, cases where the ‘events’ of a disease are not fitted in a Poisson process. If the assumption of a Poisson process for the cases of a disease is not appropriate, we have to approach our problem from a different view.

Since public health surveillance data are used in regular time intervals, the time series of the number of diseases often exhibit time dependence such as autocorrelation and seasonality. For that reason there has been a great amount of papers which are dealing with the modeling of these time series and provide forecasts of future incidence values. The possible deviations from the modeled series can be thought of as an indication of a change in the pattern of disease. Moreover, no assumption of stationarity is needed in the general surveillance setting but, if it exists the situation is simplified.

A first choice could be the Box-Jenkins (seasonal) autoregressive integrated moving average models (ARIMA) presented in Box & Jenkins (1970)[10]. Box-Jenkins models have been used in vast literature. Some studies are the following:

Choi and Thacker(1981)[23], Helfenstein(1986)[57], Nobre *et al.*(2001)[95], Reis and Mandl(2003)[106], Schnell *et al.*(1989)[124], Zaidi *et al.*(1989)[172].

Additionally, there is statistical software from which we can calculate the Box–Jenkins modeling. Such an example for statistical software is the statistical software for Public Health Surveillance (SSS1) developed by the CDC (Stroup *et al.*(1994)[148]). This software gives the ability of analyzing surveillance data, including the Box–Jenkins method.

Predictions estimate the expected incidence values and these are compared with the most recently observed disease incidence value. Several steps are necessary to proceed in the time series analysis:

❖ *Stationarity*: Speaking not strictly for a time series $\{X_t\}$ with $t = 0,\pm1,\ldots$, we can support that it is stationary if it has the same properties similar to those of the "time-shifted" series $\{X_{t+h}\}$ for each integer *h*. Now let be a time series $\{X_t\}$ with $E(X_t^2) < \infty$. The mean function of $\{X_t\}$ is :

$$\mu_X(t) = E(X_t)$$

The covariance function of $\{X_t\}$ is:

$$\gamma_X(r,s) = Cov(X_r, X_s)$$

for all integers *r* and *s*. According to these, we can give the definition of stationarity:

$\{X_t\}$ is a stationary stochastic process for all t, if the mean of the process is constant ($\mu_X(t)$ is independent of t) and the covariance between $\{X_t\}$ and $\{X_{t-1}\}$ ($\gamma_X(t,t-1)$) depends only on the time lag k.

For the case of non constant mean, traditional transformations are required to generate a stationary series from the non stationary series. Time lag differencing is used when non stationary means are encountered. For the case of the existence of dependencies between the variances and the time, square root transformations are required.

❖ *Identification and estimation*: Identification of an adequate stochastic process to describe the observed time series is needed. The tools used for identification are the autocorrelation function *(ACF)*, the

partial autocorrelation function *(PACF)* and the inverse autocorrelation function *(IACF)*.

The ACF indicates the order *(p)* of the moving average part. The PACF and IACF indicate the order *(q)* of the autoregressive part.

In the case when the orders of the process are determined, we proceed in the estimation of the parameters with the help of the maximization of a likelihood function.

❖ *Diagnostic checking*: In this case we are interested in the residuals. Residuals have to fulfill three properties:

a. The mean of the residuals should not be significantly different from zero.

b. The distribution of residuals should be normal.

c. There should be no residual autocorrelation.

An appropriate t-test for the significant difference from the zero value of the mean would be appropriate for the first property. The Kolmogorov–Smirnov test for the normality of the residuals (see Daniel (1995)[28]) and the Box–Ljung statistic (Ljung and Box (1978)[87]) can be used respectively to verify the last two properties.

When the analysis of the residuals is complete, the model can be used to forecast values and their corresponding confidence limits. The forecasts are assumed to be normal in order to calculate the 95% interval. That is the estimating value plus or minus the square root of the forecast variance.

9.1 Process Control methods and Box–Jenkins models

Williamson and Hudson(1999)[164] give a description of a combination of the Box–Jenkins models and statistical process control methods. In this paper a two-stage monitoring system is described.

An ARIMA model is developed as it is given above. Consequently, the residuals from the prediction are assumed to be approximately independent and identically distributed and they are tracked in a statistical process control. This system was performed on data from the NNDSS. In Vanbrackle and Williamson (1999)[158] this idea was further investigated and the *ARL* performance was investigated applying the Shewhart, the moving average method and the EWMA method to these residuals for four different types of shift.

Watier *et al.* (1991)[160] proposed an autoregressive integrated moving average type of model-based warning system where the alert threshold value is a function of the upper side of the prediction interval. The idea was applied to data for Salmonella in France.

Nobre and Stroup (1994)[94] used the forecast errors to calculate a probability index function to detect deviations from past observations applied to data for measles cases reported through the NNDSS.

9.2 Integer-valued Autoregressive Models

Another type of models is the integer-valued autoregressive models *(INAR)* for the analysis of time series. They have been studied theoretically by many authors (Al-Osh and Alzaid(1987)[2], Du and Li (1991)[31], Latour(1997)[79] & (1998)[80]) and an application can be found in Cardinal *et al.*(1999)[12]. This class of models is an interesting alternative to the real-valued time-series models which do not respect the nonnegative integer-valued characteristics of surveillance values. Real-valued models applied to nonnegative integer-valued observations may be an inappropriate strategy, especially for the analysis of rare events. An INAR process of order p is defined by:

$$X_t = \sum_{i=1}^{p} a_i \circ X_{t-i} + e_t$$

An epidemiologic interpretation of this formula is to consider that $\{X_t\}$ is the prevalence of the disease at time t. The prevalence at time t is the sum of

individuals remaining infected with a probability *a* in the time interval *(t−1,t)* and individuals contracting the disease in the same interval (represented by e_t).

The Steutel and van Harn's convolution operator (Steutel and van Harn (1979)[143]), denoted "$\circ$", is defined by:

$$a \circ X_t = \sum_{k=1}^{X} Y_k$$

$\{Y_k; k \in \mathrm{N}\}$ is a sequence of identically and independently distributed random variables which follow a Bernoulli distribution with parameter *a*. If we consider an integer-valued autoregressive process of order 1, the first formula can be rewritten as:

$$X_t = Y_1 + Y_2 + ... + Y_{X_{t-1}} + e_t$$

INAR models are identified using the same tools as for ARIMA models (ACF, PACF). Autoregressive parameters are estimated using either the Yule–Walker estimation technique or the conditional least-squares method.

Cardinal *et al.* (1999)[12] concluded that an INAR model provides a smaller relative prediction error than ARIMA models for meningococcal disease.

9.3 Serfling's Method

In this case Serfling (1963)[127] presented a system where a statistical analysis takes place for weekly pneumonia and influenza deaths in 108 US cities. This system is the foundation for several papers in epidemiological literature. The way it works is through a regression model which fits the non epidemic data and predicts a non epidemic level curve.

Costagliola *et al.* (1991)[26] applied Serfling's method to the French influenza-like syndrome data collected from a sentinel network from 1984 to 1988. In this paper the cases for the periods above three cases per sentinel general practitioner (SGP) were deleted. The regression equation

$$y_t = a + bt + c_1 \cos\frac{2\pi * \mathrm{t}}{52} + c_2 \sin\frac{2\pi * \mathrm{t}}{52} + c_3 \cos\frac{4\pi * \mathrm{t}}{52} + c_4 \sin\frac{4\pi * \mathrm{t}}{52} + e_t$$

is fitted in order to predict the expected non epidemic level for the following winter. In this equation y_t is the number of cases per SGP in week *t* and e_t are the residuals for which the assumption that they follow a centered normal distribution is made. The parameters are estimated by the least-squares method.

The disadvantages of this method are the limitations we have to make in order to develop an effective monitoring system. The first limitation of this approach is that we have to define at what number of cases per SGP we can consider that past observed data should be deleted when fitting the model. The second limitation is that the model assumes a seasonal period and very specific terms in the regression equation. This means that the process under study must be relatively regular over time. Thirdly, this method can be applied not in all the types of time series as it exhibits different features in terms of seasonality, number of cases, etc.

However, this approach represents a simple tool to analyze surveillance data for relatively well-known diseases as it was shown by Flahault *et al.*(1995)[34].

9.4 Log-linear Regression Model

The following model is a very interesting tool for the detection of an epidemic as it adopts for the majority of data characteristics of a statistical solution.

Farrington *et al.* (1996)[33] presented an algorithm which is constructed based in reports from Communicable Diseases Surveillance Centre (CDSC). This model is described as follows:

$$\log(\mu_i) = a + bt_i$$
$$E(X_i) = \mu_i$$
$$Var(X_i) = k\mu_i$$

The main idea, for this model, is that for each week t_i we have a baseline x_i (number of events).The assumption we make is that the x_i follow a distribution with mean μ_i and variance $k\mu_i$, where k is the dispersion parameter. Estimates are obtained by a quasi-likelihood method.

As far as we are concerned about the probable trends and seasonality of the time-series we solve these problems by fitting a linear time variable in our regression $\log(\mu_i)$ and by using observations from comparable periods in the threshold calculation(as in the *Historical Limits method* of chapter 7), respectively. This calculation involves the problem of serial correlations between baseline counts which is tackled with the absorption of their estimations in the threshold expression. The influence of baseline counts in time periods coinciding with past outbreaks is reduced by constructing weights based on adequate residuals (Davison and Snell, (1991)[29]).If we associate low weights with large residuals then our problems are solved in a great degree. This method's main disadvantage is that if we adjust it for over dispersion, then our method becomes very sensitive and as a result it detects small increases in rare diseases.

Since 1996, this method has been applied to the detection of aberrations for a set of 200–350 different types of organisms reported from laboratories. Each week, an excess score is given for each organism and if this score is higher than one, an alarm is triggered.

9.5 Other Parameter-driven Models

Zeger(1988)[173] presented another Poisson log-linear regression model as an alternative to observation-driven models described in the previous sections. In observation-driven models, $\{X_t\}$ is a function of past observations $X_{t-1,}X_{t-2,\ldots}$.

In this parameter-driven model an unobserved stochastic process generates the dependence between random variables of the process of interest. For the case of parameter-driven models, methods with linear models which are

presented by West and Harrison (1989)[161] and methods based on the Kalman filter (Kalman(1960)[66]) seem to be useful. An example for this last case is the study presented by Stroup & Thacker(1993)[144] who applied the Kalman filter to the surveillance of AIDS data.

Another application of the Kalman filter is the study of Smith and West (1983)[137]. Kidney failures with various possible changes are studied in a Bayesian framework. Representing the problem as a state space model, the multiprocess Kalman filter was used to calculate on-line posterior probabilities for the various states. Further literature can be found in Smith *et al.*(1983)[138] and Gordon and Smith(1990)[53].

In Whittaker and Fruhwirth-Schnatter(1994)[163] the same approach was used for detecting the onset of growth in bacteriological infections. An alarm was triggered if the posterior probability of a change exceeded a fixed constant. In Schlain *et al.*(1992)[122] the use of a Shewhart–CUSUM method is found, applied to recursive residuals from a continuous time first-order autoregressive model, where the parameters of the model were continuously updated by using a Kalman filter, can be found. Other examples of this approach can be found in Schlain *et al.*(1993)[123] and Stroup and Thacker(1993)[144].

Other parameter-driven models called hidden Markov models (HMMs), have been applied to the monitoring of surveillance data (Le Strat and Carrat(1999)[83], Rath *et al.*(2003)[104]) and the analysis of hospital infection data (Cooper & Lipsitch (2004)[25]). The basic idea is to associate with each X_t, an unobserved random variable S_t that determines the conditional distribution of X_t. Parameter estimations are obtained by the maximization of a likelihood function.

Some other examples of time series modeling can be found in Healy(1983)[56], Ngo *et al.*(1996)[93], Simonsen *et al.*(1997)[135] and Quenel and Dab(1998)[101].

An interesting study on this subject is the one of Cowling et al.(2006)[27] where time-series and CUSUM models are compared with the Serfling's method. It was shown that time series and CUSUM models are more effective when applied on short-term data. The study was based in the surveillance of influenza data from honk Kong and the US.

A review of several statistical approaches as well as some methodologies and time-series models, which we mentioned in this chapter with the appropriate simulations, are given in the review study of Unkel et.al.(2012)[156].

Finally, we should recommend the book of Lawson & Kleinman(2005)[82].The study of Yann Le Strat(2005)[170] in this book gives an overall review of temporal surveillance including different types of time-series models.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

PART III

Detection of Increased Rates of Incidence in a Spatial Process


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 10
Spatial Surveillance

In the previous section we studied some cases in time for the detection of an outbreak of a disease. Monitoring cases in time and analyzing temporal data, though, do not always give the statistical results epidemiologists need in order to decide the appropriate actions. There are many cases for which we construct methods for surveillance systems, from a prospective view, in order to detect a change in the data. This analysis often does not contain any information about the spatial factor. As a result, our system is ineffective since it ignores the spatial structure of the data. All surveillance methods discussed so far are examples of this. One of the main purposes of the surveillance systems that are in use is to detect changes in observed data. Ignoring the spatial structure of the data, we ignore a wide part of its foundations and so we are led to use less information.

Because of this loss, we construct suboptimal surveillance methods. An example is a local change which is smoothed (and therefore not detected) because of aggregation of the data. The case of a spatially spreading shift process is another example. These are two cases where the spatial component contains the important subject-matter information. The significance of the spatial factor varies according to the disease and the type of our problem but the exclusion of such information will severely limit not only the ability for detection but also our understanding of the process. So, further statistical analysis in a second level is needed. That is the **spatial** analysis.

Actually, most public health surveillance systems are developed with respect to the space and the time of the ‘events’. International or national organizations collect reports with the events recorded in each state, country or some specific locations. Such organizations for example might be the International Clearinghouse for Birth Defects Monitoring Systems or the Centers for Disease Control and Prevention which collects data from all over the USA. Another example is the European co-operation during the winter of 1993–1994 (Fleming and Cohen (1996)[35]) where the influenza epidemic started in

Scotland and spread south to the rest of the European countries via England and France.

In order to develop a spatial surveillance method there are two types of situation we desire to deal with. For the first situation we are dealing with various forms of level changes with possible spatial dependence between the locations of observation. The second case concerns not only level changes in the case when the data collected are spatially correlated but also the detection of changes in spatial patterns. An example of this is various forms of clustering of diseases. The case of child leukemia has been the topic in many retrospective studies.

For the temporal part there are applications for a multivariate surveillance. For example a multivariate version is applied to *the sets method*, using data on malformations from multiple sources. In that case, fixed time periods were used contrary to the univariate version, which uses the time between events. Here, the number and size of terminated sets within the time period are used. An alarm is given if each of the terminated sets is smaller than a certain number. However, as previously discussed, it would be preferable to base the surveillance system on the initial interval data. In Stroup *et al.* (1988)[145] the possibility of using multiple time series for the detection of excess deaths from pneumonia and influenza was discussed. Here, one-step-ahead forecasts are used.

Constructing surveillance methods for spatial processes is a complicated problem. For the spatial analysis we consider different assumptions for the observed process and different ways of observing and modeling this process. In the case of spatial surveillance a change in a parameter of the distribution of the observations might have a clear spatial interpretation, e.g. a stronger tendency for clustering. In Lawson (2001)[81] a discussion of how to generalize various kinds of spatiotemporal models to allow for prospective surveillance is given. Various problems when applying surveillance methods to spatial public health situations were pointed out.

In the study of Kulldorff (2001)[72], the issue of the prospective monitoring of clusters is also presented by using a modification of the spatial scan statistic proposed in a previous study of his (Kulldorff (1997)[71]). The new statistic is a combination of the spatial and temporal features. The

spatiotemporal scan statistic is used prospectively for fixed time windows as it is combined with the cumulative sum methods.

Dealing with spatial surveillance means the acceptance of the assumption that there is a spatial model for the observed data. Another approach based on this spatial model is proposed by Järpe (1999)[59]. This approach was used in the surveillance of clustering in a spatial log-linear model with a fixed lattice: *the Ising model*. In this case we proceed in the sufficient reduction of the spatial structure in order to tackle with the problem of the loss of information based on the spatial factor. The result of the proper reduction is a univariate statistic involving the sufficient spatial factors for each time point. A complete separation of the spatial and the temporal factors was possible.

The same purpose of the reduction of the spatial structure is studied through the method of the likelihood ratio by Järpe (2000)[60]. The main subject is to deal with a shift in the process with respect to the spatial factor as time increased. A likelihood ratio statistic which involves a sufficient reduction in the spatial structure is proposed. In this case, though, a complete separation of the spatial and temporal components was not possible owing to the nature of the problem. Different ways of treating the multivariate structure in the spatial surveillance situation was discussed. As an application, the problem of an increased rate of radiation was investigated. An evaluation and comparison with the system that is currently in use in Sweden, which is based on a moving window, was made. The situation with a spreading shift process would correspond well to the surveillance of influenza, where the disease spread across Europe from Scotland (Fleming and Cohen (1996)[35]).

Several issues are analyzed in the studies of Vallet and Lawson (2011) [157], Assunção and Correa(2009)[6] and Gallego(2010)[49] for the purpose of the prospective timely detection of incident disease clusters in space and time.

In public health, this idea is widely spread the last few years since it considers not only the temporal feature of a case as we saw in the parts I & II. The space factor is proved to be very important in monitoring the public health. That is the reason why, several approaches have been developed as we described them above.

Kulldorff(2005)[73] presented a very good study in Lawson and Kleinman's(2005)[82] book where he describes the role of the scan statistics for

geographical diseases. He uses mostly Poisson or Bernoulli distributed data and shows how these scan statistics are implemented in some maps or methods such as the CUSUM scheme. Moreover he gives some ideas of proceeding in geographical clusters of the disease monitored.

A spatiotemporal extension of the spatial scan statistic appears in Kulldorff et al.(2005)[74]. Different ways of constructing space–time scan statistics based on surveillance theory is presented in the study of Sonesson (2007)[139]. Sonesson showed how spatial-temporal scan statistics can be embedded in a CUSUM framework and applied these methods to the detection of an increased rate of Tularemia in Sweden. The same approach was used by Marshall et al.(2007)[89] in order to construct a CUSUM method for monitoring the local Knox statistic tests for space and time clustering each time there exists a new observation.

A space-time scan statistic for the detection of an outbreak in public health is presented in Takahashi et al.(2008)[150]. An improved spatial system is presented in Johnson(2008)[63] for the case of the West Nile virus in U.S. through generalized linear mixed models. The local Knox statistic is used.

Some prospective scan based methods are reviewed in Woodall et al. (2008)[168]. Issues that are related to the spatio-temporal scan based statistics' evaluation are referred.

Tsui et al.(2011)[154] developed a general framework for spatial and spatiotemporal surveillance based on likelihood ratio statistics. The CUSUM scheme and Shiryayev-Roberts statistics are special cases under such a general framework.

An application of the Poisson CUSUM in rare events from different regions was developed by Rogerson and Yamada (2004a)[117]. The purpose of this paper is to construct a multiregional surveillance system with the help of a Poisson CUSUM which is applied in infrequently appeared counts from different locations. On the CUSUM scheme for the spatial case it was also presented by Rogerson and Yamada (2004b)[118], a comparison between univariate and multivariate CUSUM approaches. Several multivariate CUSUM schemes are given in the paper of Jiang et.al.(2011)[61]. The multivariate case gives us thoughts for further applications in the future for the spatial case.

Spatio-temporal surveillance is the objective of Rodeiro and Lawson(2006)[111]. Several methodologies are discussed on this paper in order to construct an effective surveillance system with spatial features included. These methodologies are based on hierarchical space-time models.

A review for spatial, and spatial–temporal systems that can be used to facilitate the early detection of infectious disease outbreaks is given in Chen et.al.(2011)[14].

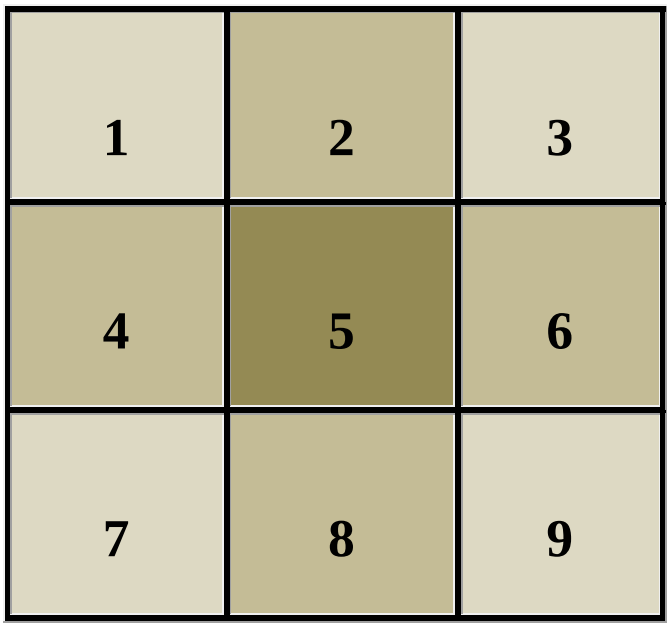
A review of the available software for spatio-temporal surveillance of diseases is given in Robertson and Nelson(2010)[110].

A general review and interesting issues for spatial clustering and space-time scan statistics are given for further study in the book of Tango(2010)[152].

In the following sections we are going to examine the usage of the cumulative sum methods in spatial surveillance. We focus on data from different regions (neighboring or not) and we present how the CUSUM applies to public health problems. Data may be normally distributed but it is common to use the Poisson model too. We also present the variety of problems coming up as someone examines this type of monitoring in public health.

10.1 Cumulative Sum Methods and Spatial Surveillance

In this section we are going to see the CUSUM method applied in a region. If we want to take into account all the available information generated from the spatial factor, then we should divide our region of interest. In this section we divide the region in 9 locations/areas. A graphical image of our location is given below:



From the projection of our area illustrated above, we assume that we are interested in a squared region 3x3 which is divided in 9 smaller squared locations of equal size 1x1. In a few words we generate normally distributed data for each one of these areas with a mean of zero and standard deviation of one. We generate 20 values for each region and we make the assumption that these 20 values correspond to different periods (time).

With different colors we want to present the different rate of change in the mean of the different areas. The changes are described in the following table:

Table 10.1: The table with the changes in the mean for each region

Regions	***Change in the Mean***
1,3,7,9	+ 0.3
2,4,6,8	+ 0.5
5	+ 1

For each region we also make the assumption that it maintains its CUSUM scheme. We take a value of $k=1/2$ and we want to figure out the limit h for a fixed value of ARL_0 (say $ARL_0 = 120$). Then from the equation *(6.2)* we have:

$$h \approx \left(\frac{ARL_0 + 4}{ARL_0 + 2}\right)\ln\left(\frac{ARL_0}{2} + 1\right) - 1.166 \approx 1.0164 * \ln 61 - 1.166 \approx 3.01$$

The generated values before and after the change in the mean for each one of the nine regions are presented in the tables of the *Appendix E.1*. We proceed in calculating the statistics of the CUSUM scheme for each of the nine locations and we have the CUSUM graphs respectively in *Appendix E.2*.

We conclude that for the small change in the mean (+0.3) (areas 1, 3, 7, 9) we have the longest delay in realizing the alarm (region 1: 9 time periods, region 9: 10 time periods). Additionally, we have a false alarm for the 3rd area and no alarm at all for the 7th area.

For the change in the mean of +0.5 we have a time delay of 5 periods for the areas 2, 6, 8 and for the 4th location we have a time delay of 8 periods.

For the biggest change in the mean (+1.0) we have the smallest delay (region 5: 3 time periods).

10.2 Problems

These nine systems for each region give a unique CUSUM scheme and we can realize which area has an outbreak or not. However this thought most of the times do not give us reliable conclusions. That is for two main reasons; the unrealistic limit *h* and the small changes in the mean of a disease in some areas.

10.2.1 The problem of limit h

If we have a fixed average run length until the first alarm of about *120*, we have to construct a more reliable and a more realistic limit. The value of *ARL* practically *means that for each region we should have an average time of 120 periods before our system trigger an alarm.* Thus, the value of the limit *h=3.01* is too low and we have as a result frequent alarms that may be neither true nor reliable.

If we want to have a CUSUM scheme for each one of the nine regions we have to adjust our threshold which we are going to use in each case. Adjusting this threshold, leads us to use the number of areas in which the location of interest is divided. That is *m=9*. If we have a fixed $ARL = 120$ and *k=1/2* with the purpose of maintaining all the regional charts, then for all 9 areas we use the approximation given in Raubertas(1989)[105]:

$$ARL_0 = \left[1 - \left(1 - \frac{1}{ARL_0^*}\right)^{1/m}\right]^{-1} \quad (10.1)$$

This is based on the fact that the *ARL* follows an exponential distribution approximately. Thus for the new ARL_0 we have the following:

$$ARL_0 = \left[1 - \left(1 - \frac{1}{ARL_0^*}\right)^{1/m}\right]^{-1} = \left[1 - \left(1 - \frac{1}{120}\right)^{1/9}\right]^{-1} = 1075.99$$

Then the new *h* is given as follows from the *(6.2)* equation:

$$h \approx \left(\frac{ARL_0 + 4}{ARL_0 + 2}\right)\ln\left(\frac{ARL_0}{2} + 1\right) - 1.166 \approx 1.00185 * 6.2897 - 1.166 \approx 5.135$$

In a few words, with this value of ARL_0 and therefore the new value of *h* we achieve that the average time until the first alarm (false alarm) over the *m* regions is equal to *ARL*. For these values we proceed in our simulation using the data which were generated before, for the 9 locations. We construct the CUSUMs for each of the nine locations with the new limits. The appropriate graphs are given in the *Appendix E.3*.

From the simulation, we notice that now we have a more realistic but pessimistic limit and as a result it is difficult for the system to detect small changes of the mean. The areas 3, 7, 9(with the smallest increase in the mean) this time do not trigger an alarm. Additionally, at region 3 where before we had

a false alarm now we do not have an alarm at all. From the regions with the small increase in the mean only the 1st triggers an alarm (region 1: delay of 12 time periods).

For the group of areas with a normal increase of +0.5 in the mean, only the 8th area does not trigger an alarm. The others trigger an alarm but there is a large delay especially for the 4th location (alarm at the 40th observation).

The fifth's area CUSUM with the largest increase in the mean gives an alarm with a delay of 8 time periods.

Generally, four regions do not trigger an alarm and five trigger an alarm. Comparing these results with the CUSUMs of the limit *h=3.01*, we see that the alarms are fewer. In the case of the *3.01* limit, only one area did not trigger an alarm. Moreover, in that case we had a false alarm too. Using the approximation of the limit *h=5.135*, we improve our system from the view of not having a false alarm and from the perspective that the average time until the first alarm over the set of the 9 charts, is equal to *ARL*. From this point of view, this is a more realistic limit and we are confident for our decisions if we have an alarm. The absence of false alarms makes our system more reliable. On the other hand, this value of the limit *h* is absolutely conservative and leads both to the absence of alarms and to larger time-delays in realizing the change in the mean.

10.2.2 The problem of small changes

It is understandable that epidemiologists could miss an outbreak (i.e. of a disease) which takes place in a lot of regions. As we saw in the simulation above, small changes in each region are hardly detected and even if regional alarms are triggered this is usually done with a large time delay. Thus, small changes in the rate of a disease are a problem for each region's system. That is why a more overall system has to be constructed. A small change in one region may not give us an alarm but if small changes take place in the neighbor regions, then it is easier for the total system to detect the change and trigger an alarm. There are two ways to solve this problem and proceed in the construction of a proper system:

❖ The first is to ***maintain the CUSUM charts for local Neighborhoods around each region.***

Raubertas (1989)[105] suggested the CUSUM method not only for each location but also for its surrounding neighborhood. That is the main idea of constructing the appropriate statistics for the appropriate locations. In other words, we use some *weights* based on the distance between the locations. These weights are increasing, if the distance is small and decreasing if the distance is large.

Our purpose is to detect a change in the mean of one or more regions through local statistics. An immediate detection of a shift from the null hypothesis (where there is no spatial pattern and all regions have zero means) to the situation where one or more regions witness an outbreak of the event is of interest. Rogerson(2005)[114] at his study in the book of Lawson and Kleinman(2005)[82] gave us a detailed presentation of how this method works.

At each location, we construct a local statistic, by using a Gaussian kernel, represented by a weighted sum of the regional values:

$$y_{it} = \sum_j w_{ij} x_{jt} \quad (10.2)$$

Where w_{ij} are the *weights* and are expressed as follows:

$$w_{ij} = \left(\sqrt{\pi}\sigma\right)^{-1} \exp\left\{-\frac{d_{ij}^2}{2\sigma^2}\right\} \quad (10.3)$$

where σ is the width of the Gaussian kernel (chosen to coincide with the likely size of any emergent spatial cluster), and d_{ij} is the distance from the centre of region *i* to the centre of region *j*.

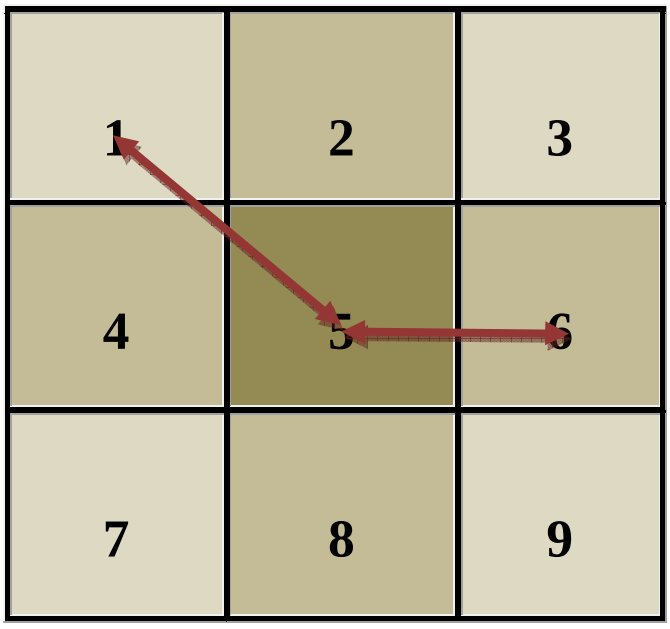
An issue for this methodology is the case of the locations near edges where there are not as many neighboring locations as in other regions. As a result, the sum of the squared weights $\sum_j w_{ij}^2$, and the variance of the

local statistic (which is based on the sum of the squared weights), will be smaller for regions near edges than for other regions which have more neighbor regions.

Giving solution to this problem, modified scaled weights are used in order to have equal variances for all the statistics of the different regions. Modified weights are expressed by the following equation:

$$w_{ij}' = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}} \quad (10.4)$$

Using the normally generated data from the previous simulation of the 9 regions, the centre of our area of interest is the fifth location. In order to find the appropriate weights for this region we have to figure out the distances. We presented above the squared region of interest which is divided in 9 areas. We make the assumption that the total area is of dimension 3x3(e.g. in kilometers). For our convenience, we proceed in the example with the scale of kilometers (and without transform the distances in meters). We also assume that the 9 locations are squared areas with equal dimensions. Therefore each area is of dimension 1x1. The distances are calculated considering the centers of each area.



Thus, we have the following table with the distances d_{ij} for the fifth location:

Table 10.2: The distances of the 9 regions from the 5th region

d_{ij}	***j***								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
i	**5**	$\sqrt{2}$	1	$\sqrt{2}$	1	0	1	$\sqrt{2}$	1	$\sqrt{2}$

The weights for this area are calculated for $\sigma = 1$. For the corner areas 1,3,7,9 and the areas 2,4,6,8 the weights are equal since the distances are the same. For the areas 2,4,6,8 the weights are given below from *(10.3)*:

$$w_{52} = w_{54} = w_{56} = w_{58} = \frac{1}{\sqrt{3{,}14}} \exp\left\{-\frac{1}{2}\right\} = 0.3422$$

For the areas 1,3,7,9 and 5 the weights are calculated with the adjustment (from *10.4* equation) we mentioned above:

$$w_{51} = w_{53} = w_{57} = w_{59} = \frac{1}{\sqrt{3{,}14}} \exp\left\{-\frac{2}{2}\right\} = 0.2076$$

$$w_{55} = \frac{1}{\sqrt{3{,}14}} \exp\{0\} = 0.5643$$

Then we calculate the sum of the squared weights:

$$\sum_{j=1}^{9} w_{5j}^2 = w_{51}^2 + w_{52}^2 + \ldots + w_{59}^2 = 4 * w_{51}^2 + 4 * w_{52}^2 + w_{55}^2 \Leftrightarrow$$

$$\Leftrightarrow \sum_{j=1}^{9} w_{5j}^2 = 0.1724 + 0.4684 + 0.3184 = 0.9592$$

So, using the adjustment for the corner areas we have:

$$w_{51}' = w_{53}' = w_{57}' = w_{59}' = \frac{w_{51}}{\sqrt{\sum_j w_{5j}^2}} = \frac{0.2076}{\sqrt{0.9592}} = 0.2119$$

And for the centre area (5^{th} location) we have:

$$w_{55}' = \frac{w_{55}}{\sqrt{\sum_j w_{5j}^2}} = \frac{0.5643}{\sqrt{0.9592}} = 0.576$$

So, for these weights we calculate the statistic y_{it} for the (centre) area 5 (from (*10.2*) equation). The new values from which we calculate the CUSUM statistic are given below:

Table 10.3: The transformed observations of the 5^{th} area after we include the weights

Time Period	***Values of area 5 with the Weights of Neighbor regions***
1	-0,47538
2	-0,55773
3	1,22710
4	0,07724
5	-1,48427
6	-0,52506
7	1,39312
8	-1,87274
9	0,80134
10	0,98338
11	0,33103
12	0,74165
13	0,74571
14	0,46018
15	0,88657

16	0,51832
17	0,02801
18	-0,45050
19	-1,31354
20	-1,62783
21	2,75034
22	0,72749
23	2,54517
24	1,45112
25	2,57387
26	1,26622
27	3,08814
28	2,42695
29	2,05999
30	2,46690
31	2,01955
32	1,91117
33	-0,41139
34	-0,17642
35	1,55448
36	1,24310
37	0,89420
38	2,09192
39	0,70456
40	1,15628

For the value of *h* to be *3.01* we have an alarm at the 23rd observation (3 time periods delay). Nothing changed in the time of the alarm, compared with the system without the weights of the neighbor regions of the 5th area.

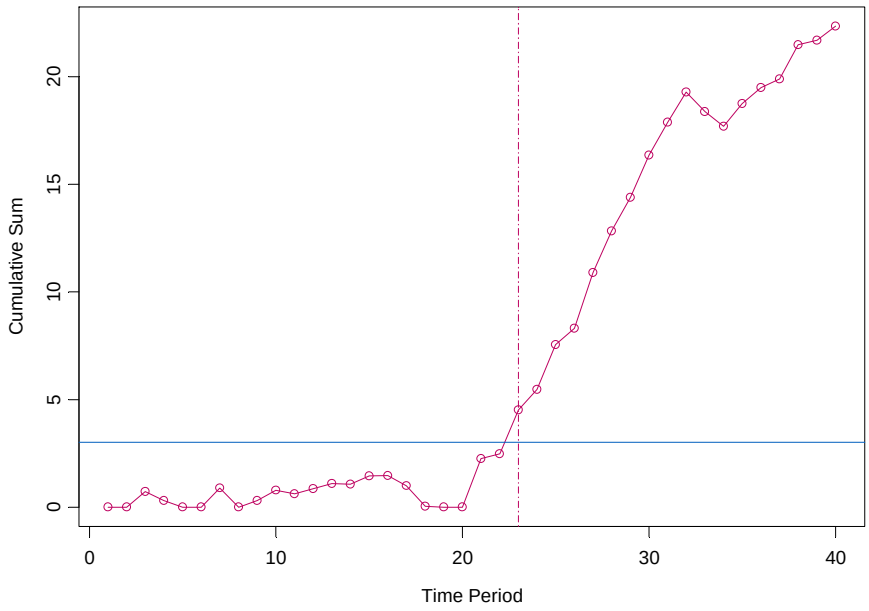


Figure 10.1: The CUSUM scheme for the weighted observations of the 5th region for a limit h=3.01.Alarm at the 23rd observation

With the more appropriate value of *h* but a more conservative approach we have an alarm at the 24th observation. For this value of *h* though, we see great improvement using the weights. Without the weights, we had an alarm at the 28th observation and now we have an alarm at the 24th observation for area 5. ***The delay has decreased and with this way we manage to transform a conservative approach of the system (of the limit h=5.135) to a monitoring system with a short time delay.***

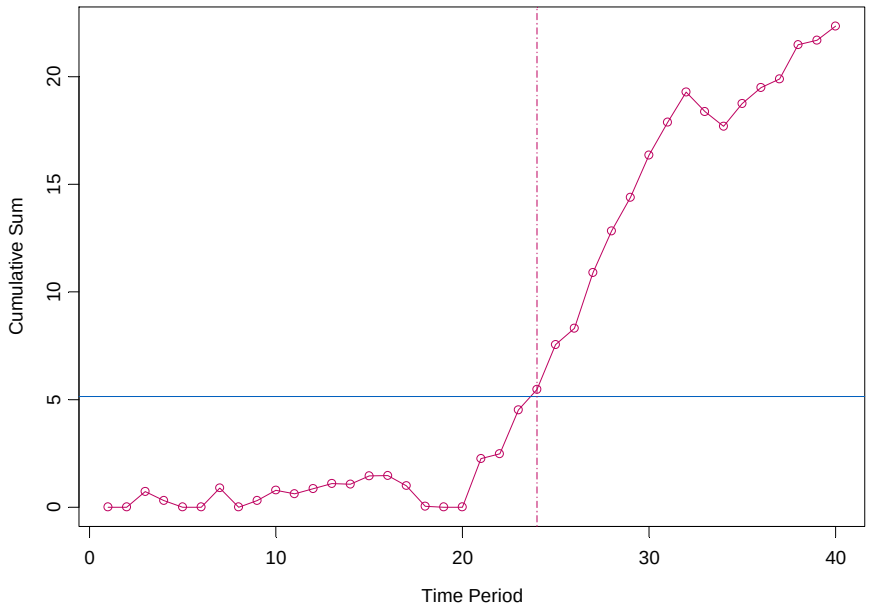


Figure 10.2: The CUSUM scheme for the weighted observations of the 5th region for a limit h=5.135.Alarm at the 24th observation

For spatial disease surveillance with data from several locations, Raubertas (1989)[105] suggested the ***Poisson CUSUM method*** for each of these locations. To account for the positive spatial correlation between nearby locations, the author suggested pooling within neighborhood observations, using closeness as weight. Making this modification the sensitivity of the Poisson CUSUM is improved. We have an alarm as soon as one of the CUSUM systems of a location signals an alarm. The measures of ARL_0 and ARL_1 are proposed for the system's evaluation. For Poisson variables, one can monitor the y_{it} :

$$y_{it} = \sum_j w_{ij} x_{jt} ,$$

where the x_{jt} is the observed count in region j at time t, and w_{ij} is a weight associated with the distance from region i to region j.

These observed quantities are then compared with their corresponding expectations, $\sum_j w_{ij}\lambda_{0,jt}$ for region *j* at time *t*, and used in a CUSUM for region *i*.

For the issue of the thresholds, Monte Carlo simulation of the null hypothesis may be used as it is referred in Rogerson(2005)[114]. Observed counts are realizations from Poisson or Normal distributions with parameters set equal to the corresponding expectations. The thresholds should be determined using $sARL_0$ with a desired average run length of ARL_0, where $s < m$. The value of *s* is determined via simulation with the purpose of achieving the desired average run length. The greater the correlation between the local regional statistics, the lower *s* will be relative to *m* (Rogerson and Yamada (2004a)[117]).

❖ The second solution is to develop ***a global spatial statistic within a CUSUM chart***. In spatial surveillance we are facing problems involving both spatial and temporal components. To deal with these problems we can use different approaches. One example is the Rogerson's approach. According to this approach we proceed in the surveillance in time of a purely spatial statistic which describes the spatial pattern for each time point. This is the case when using a univariate test statistic designed for a retrospective test and following it through time by using a surveillance method. This approach was used in Rogerson (1997)[112], where a modification of the retrospective test suggested in Tango (1995)[151] for both, general and focused clustering, was used prospectively within a CUSUM method. Rogerson used Tango's statistic for a general test clustering:

$$C_G = (r - p)' A (r - p) \quad (10.5)$$

where *r* and *p* are *mx1* vectors containing the observed and expected proportions of cases in the *m* regions of interest. *A* is a *mxm* matrix containing elements a_{ij} that measure the closeness of region *i* to region *j*.

The function of a_{ij} according to Tango is expressed with respect to the distance, d_{ij} between two regions:

$$a_{ij} = \exp\left\{-\frac{d_{ij}}{\tau}\right\}, i \neq j \quad (10.6)$$

with $a_{ii} = 1$. The scale parameter τ is related to the size of the cluster; larger values of τ are effective in detecting larger clusters. Tango uses $\tau=5$ .The test statistic relies on successive computations of Tango's statistic as new observations become available.

Tango's test, like the most of the tests for the detection of clusters, is used in a retrospective way. We are interested, though, in monitoring the diseases in some locations. That actually means we care for prospective surveillance. The main idea in this prospective case is to detect emerging clusters and minimizing the delay of their occurrence, simultaneously.

It is not appropriate the fact of using Tango's statistic after each new observation. That is because we test multiple hypotheses. For that reason, we calculate the expected value and variance of the Tango statistic after the next observation, conditional upon the current value of the statistic. The expected value and the variance are used to transform the Tango statistic that is observed after the next observation into a z-score. These z-scores are used in the CUSUM scheme. The CUSUM approach developed in this paper gives us the needed prospective view and at the same time solves the problem mentioned above. For the evaluation of the system, the measures of ARL_0, ARL_1 and the Median Run Length were used.

Rogerson and Sun (2001)[116] show how a similar approach may be used to monitor changes in the nearest neighbor statistic.

The same approach was used in Rogerson (2001)[113]. In this case the purpose of our method is to combine the Knox statistic suggested in Knox (1964)[70], for space and time interactions with cumulative sum methods. Rogerson developed a local version of the Knox statistic and

presents its usage in a retrospective way, in order to identify the particular observations that are associated with space-time interactions. Then the local Knox test is used to demonstrate how the Knox statistic can be mixed with the CUSUM methods for the purpose of the online monitoring of probable changes in space-time interactions as new data are collected. With such a way, a simple retrospective system is transformed in a system with a prospective view.


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Chapter 11
Conclusions and Future Research

For the purpose of this study we were influenced by the study of Sonesson and Bock(2003)[140] who presented a review of public health surveillance techniques using the SPC. We developed the methods described in this paper and we made some simulations for the extraction of further conclusions on these issues. Additionally, in each case we presented further studies for the period 2003-2012 which may guide the reader for further research in the future.

Constructing a surveillance system for public health is a very complex process which requires taking into consideration several factors such as the seasonality, the frequency of appearance, the trend, the spatial and the temporal approach of a disease etc. It is essential to make the appropriate decisions for the appropriate type of disease and proceed in the development of a monitoring system according to several features of this disease.

An outbreak of rare diseases is the object of study of several papers since this case is difficult to detect. That is because even the slightest increase in the rate of a disease might mean an outbreak. The smallest the baseline rate of a disease the hardest our system is to detect an outbreak. That is the case of several comparisons between different methods and which allow us to study their behavior. Using rare diseases (the most difficult type of diseases to detect an outbreak) gives us an extra motive since their strict assumptions and their small rate of occurrence make the detection of a change more difficult and therefore we have the most appropriate type of disease for a comparison between methods.

In order to compare methods we may use the measures of evaluation we mentioned in the 3rd chapter of this study. In terms of the average run length, we want a small run length when an alarm has occurred and a large run length when our system triggers a false alarm. Therefore, the smallest the probability of false alarm and the highest probability of successful detection, the better will be a

method. The same idea is given for the measure of the predictive value where we want as high a value as possible. The trust in the alarm is a very sensitive issue in the public health surveillance and so we would like to have the largest possible value for the predictive value. In terms of the expected delay to detect a change, it is rational to think that we should have a small value if we want an effective monitoring system. Especially in public health, where the delay is an important factor which may determine our preventive actions, this measure is used regularly.

In our study we gave an example of a comparison between the sets method and the Cuscore where we calculated the expected time delay. In this comparison, we came to the conclusion that the Cuscore is more effective for rates of a disease up to 5 ‘events’ per 10000 individuals.

In the most cases in public health, though, we need as much information we may derive from the data as possible. Combining several evaluation methods in public health surveillance is of vital importance. When we are dealing with human lives we need to present as much data as possible. Thus, there is a strong need to compute as much measures of evaluation as possible for a method. In theory, such need does not exist, but the sensitivity of such a field demands from us, in practice, the construction of the perfect surveillance system. For example, in the study of Frisén and Wessman (1999)[48] it was used the property of the *constant Predictive Value to be fulfilled for the SR method,* for the case of a change in the mean of a normal distribution. It could be expected that this is the case also for a shift in a Poisson process but this remains to be verified.

The predictive value as a measure actually presents us the balance between the false alarms and the time delay. The knowledge of this measure is very useful for us and provides us with important information for the system’s behavior. Especially in the field of epidemiology, this would be useful as the investigators should not ignore an alarm since an alarm could be interpreted and provide significant information whether it is too late or too early.

Beyond the measures of evaluation of a method, we faced the issue of the optimality. Several approaches are constructed in a suboptimal way. Methods such as the sets method are suboptimal ones. On the other hand, the likelihood ratio method, which is an evolution of the Shiryaev-Roberts method, is an

optimal one in terms of the minimum expected delay of realizing an alarm for a fixed probability of a false alarm. We presented the Poisson case for the *LR* and *SR* methods and we conclude that they are preferable to use them in our Poisson issues because they give optimal results. Using such methods in order to construct some new statistics is a continuous challenge.

Another issue, is the popular Cumulative schemes. Especially, when we are taking into account the small incidence rate, the usage of such schemes is something standard even for comparison reasons (such as in the study of Rolfhamre and Ekdahl(2006)[119]). The Poisson CUSUM, Lucas suggested, is used by more and more scientists in order to observe counts. However, several modifications can be made. Such a modification might be the Bernoulli CUSUM which is referred in several recent studies and offers a field of further investigation.

A new modification of the CUSUM is mentioned in Shu et al.(2010)[131]. That is the weighted CUSUM according to the time the observations were recorded. These weights are stimulating; especially, when we make the thought of the continuous change in the population size. Instead of assuming a fixed change in the population size (which in theory is a good assumption but not in practice) it would be of interest to apply this idea in practice where the population size is random. The CUSUM scheme is of great interest among the scientists and we have the strong belief that this scheme and its modifications will be the concern of many studies in the future.

An interesting study for the general field of SPC is the paper of Wu et al.(2009)[169].He proposed a control chart based on the cases we mentioned in chapters 5 & 6. In order to construct a new chart, the cases of the ‘time between events’ and the ‘number of events’ were combined. A new control chart was built which takes into consideration the rate of the number of events to the time they occurred. It could be interesting to apply such a case in the field of public health. This remains to be studied in the future.

In order to develop an effective surveillance system we have to rely on the appropriate data. Epidemiologists are interested in forecasting and giving a reliable estimation about a possible or a probable outbreak. To construct such a good monitoring system is not easy, especially when we are referred to the case where the Poisson assumption is not appropriate. In that case we have to take

into account some important factors such as trend, seasonality and the correlation between random variables of the stochastic process. Decision mechanisms are similar but the methods vary among different types of disease. For example it is almost impossible to develop a surveillance method which is as much effective for the common diseases and rare diseases as well. If we construct a method based on a common disease, a detection of an outbreak in a rare disease is something relatively difficult. The same thing applies for the exact opposite case. A method based on irregularly appeared (rare) disease is a very sensitive method for a common disease and gives frequent false alarms. For more issues on this subject we strongly recommend the study of Yann Le Strat (2005)[170] which is included in the book of Lawson and Kleinman(2005)[82].

We also remind the measures of specificity and sensitivity which are commonly used in order to compare different time surveillance series. We can conclude that since each disease and each surveillance system has its own characteristics and features, detection of an outbreak is not relied so much on the statistical methods such as on the characteristics of the system. These might be the quality of the data collected, the stability of the reporting mechanism over time, the reporting delays etc. If the surveillance system is not effective, it is easy to produce false results. So, it is vital for us, first of all, to require a good knowledge of the surveillance system before analyzing the data. Another important factor beyond statistical analysis is the epidemiologist's experience. The more their experience the more effectiveness of the system and the better the decisions in each time frame.

In all the second part of this study we do not include the spatial factor of our data. For a fixed time frame when there is a significant difference between the expected and the observed number of cases, it is reasonable to have a look at the spatial distribution of the cases in order to check for the existence or absence of localized clusters. Examining the spatial surveillance in the third part of our study, we prove through simulations the significance of this factor as well as its usefulness to the epidemiologists' decisions.

We also saw through our simulations how to construct a method with a spatial interest and how the CUSUM scheme is applied in this section. We also mentioned how to improve our CUSUM schemes with a more appropriate limit

h for an area using the distances of the neighbored regions as weights. Without the weights the centered area (region 5) gave us an alarm at the 28th observation and using the weights we had an improvement of about 4 time periods in realizing the alarm. In such a way we developed a system with a more appropriate but also a more conservative limit *h* and which gives us an alarm with a slight difference from the strict value of the limit (*h=3.01*).

Several issues can be covered by spatial surveillance such as the map design, the geographical clustering etc. We strongly believe that the tendency in epidemiology is the spatial-temporal surveillance since several studies and books in the last few years have been published. For further information on these issues the reader may search for the book of Lawson and Kleinman(2005)[82]. Additionally we can mention the book of Waller L.A and Gotway C.A. (2004)[159] where some interesting complex and advanced issues of statistical analysis of spatial data are given such as the function of the maps of analyzing patterns and clusters as well as several applications of models to spatial data.

Additionally, the study of Tsui et al.(2011)[154] gives us an interesting example which may guide future studies. In this study a general framework is given for spatial and spatiotemporal surveillance based on likelihood ratio statistics. Furthermore, it is shown that the cumulative sum (CUSUM) and Shiryaev-Roberts statistics are special cases under such a general framework. There are several methods initially developed for temporal surveillance but in order to have a more objective view of the surveillance, we could develop these statistics in order to work for spatial and spatiotemporal surveillance. Especially for the case of the LR method (and its modifications) which provides us with optimal results, enhanced statistics with a spatial frame included, could be constructed. This is a field which demands further research.

Another example on this case could be the study of Rogerson and Yamada(2004a)[117] who applied the Poisson CUSUM into a multiregional surveillance system. The spatial case of surveillance demands the conception of the multivariate idea. Thus, it could be interesting the construction of multivariate methods in order to develop more accurate and appropriate monitoring systems. Such examples are given in Rogerson and Yamada(2004b)[118] and Schiöler and Frisén(2012)[121].The multivariate case

of spatial surveillance is the main idea of several recent studies and further research in the future is necessary on this field.

Finally, a general review in the surveillance of public health as well as a guidance for future research is given in the papers of the Tsui et al. (2008)[155], Frisén(2011)[42], Woodall(2006)[167] and Unkel et al.(2012)[156]. Also, general concepts of monitoring the public health are mentioned in the study of Han(2010)[54]. These studies are highly recommended for future research.

Appendix A

The code (we used in S-plus) for the simulation of the Sets method is given below:

-A.1-

```
####### Code for a Small Scale Area #######

#k function from (5.9)#
> k<-function(g){
+ 4.61/g}
> k(7)
[1] 0.6585714

#normal rates#
> d<-c(1:10)
> p0<-d/10000
> p0
 [1] 1e-04 2e-04 3e-04 4e-04 5e-04 6e-04 7e-04 8e-04
9e-04 1e-03

#expected size of the set i from (5.1)#
> c0<-(1-p0)/p0
> c0
 [1]  9999.000  4999.000  3332.333  2499.000  1999.000
1665.667 1427.571 1249.000 1110.111  999.000

#number of false alarms#
> r<-1

#total number of births#
> b<-400*12*20
> b
[1] 96000

#The size of a sequence of sets n from (5.10)#
> n<-function(P){
+ log(P)/log(1-exp(-k(7)))
+ }

#Calculating the proper number of n from (5.10) and
(5.14)#
> P1<-function(n){
+ r/(b*p0[1]-n+1)}
> P1(3)
```

```
[1] 0.1315789
> n(P1(3))
[1] 2.782244####### Thus n=3 ########

----------------------------

> P2<-function(n){
+ + r/(b*p0[2]-n+1)}
> P2(3)
[1] 0.05813953
> n(P2(3))
[1] 3.902689
> P2(4)
[1] 0.0617284
> n(P2(4))
[1] 3.82052 ######## Thus n=4 #####

----------------------------

> P3<-function(n){
+  r/(b*p0[3]-n+1)}
> P3(4)
[1] 0.03875969
> n(P3(4))
[1] 4.458912 ###### Thus n=4 ######

------------------------

> P4<-function(n){
+  r/(b*p0[4]-n+1)}
> P4(5)
[1] 0.02906977
> n(P4(5))
[1] 4.853558 ###### Thus n=5 ######

-----------------------------

> P5<-function(n){
+  r/(b*p0[5]-n+1)}
> P5(5)
[1] 0.02272727
> n(P5(5))
[1] 5.191207 ###### Thus n=5 ######

---------------------------------

> P6<-function(n){
+  r/(b*p0[6]-n+1)}
> P6(5)
[1] 0.01865672
> n(P6(5))
```

```
[1] 5.461948 ##### Thus n=5 ######

------------------------------

> P7<-function(n){
+  r/(b*p0[7]-n+1)}
> P7(6)
[1] 0.01607717
> n(P7(6))
[1] 5.666082 ##### Thus n=6 #####

--------------------------------

> P8<-function(n){
+  r/(b*p0[8]-n+1)}
> P8(6)
[1] 0.01392758
> n(P8(6))
[1] 5.862978  ###### Thus n=6 #####


------------------------------

> P9<-function(n){
+  r/(b*p0[9]-n+1)}
> P9(6)
[1] 0.01228501
> n(P9(6))
[1] 6.035128  #### Thus n=6 #####

--------------------

> P10<-function(n){
+  r/(b*p0[10]-n+1)}
> P10(6)
[1] 0.01098901
> n(P10(6))
[1] 6.188064   ##### Thus n=6 #####

----------------------------------

#####The table with the proper n and P #####
> a<-matrix(c(0),10,2)
> a[1,1]<-3
> a[1,2]<-P1(3)
> a[2,1]<-4
> a[2,2]<-P2(4)
> a[3,1]<-4
> a[3,2]<-P3(4)
> a[4,1]<-5
> a[4,2]<-P4(5)
```

```
> a[5,1]<-5
> a[5,2]<-P5(5)
> a[6,1]<-5
> a[6,2]<-P6(5)
> a[7,1]<-6
> a[7,2]<-P7(6)
> a[8,1]<-6
> a[8,2]<-P8(6)
> a[9,1]<-6
> a[9,2]<-P9(6)
> a[10,1]<-6
> a[10,2]<-P10(6)
> a
      [,1]       [,2]
 [1,]    3 0.13157895
 [2,]    4 0.06172840
 [3,]    4 0.03875969
 [4,]    5 0.02906977
 [5,]    5 0.02272727
 [6,]    5 0.01865672
 [7,]    6 0.01607717
 [8,]    6 0.01392758
 [9,]    6 0.01228501
[10,]    6 0.01098901
```

The n

```
> n<-c(a[,1])
> n
 [1] 3 4 4 5 5 5 6 6 6 6
```

The P

```
> P<-c(a[,2])
> P
 [1]   0.13157895   0.06172840   0.03875969   0.02906977
0.02272727 0.01865672 0.01607717 0.01392758 0.01228501
0.01098901
```

####Calculating the number of the expected births in a sequence which signals an alarm after the increase from (5.15)####

```
> e<-c0*n/7
> e
 [1] 4285.2857 2856.5714 1904.1905 1785.0000 1427.8571
1189.7619 1223.6327 1070.5714  951.5238  856.2857
> q<-1-(4.62/(exp(4.62)-1))
> a1<-e*q
> a1
 [1] 4088.2788 2725.2466 1816.6494 1702.9384 1362.2144
1135.0651 1167.3787 1021.3542  907.7795  816.9198
```

####Calculating the expected time duration in which the a1 infants are born####

```
> t1<-a1/400
> t1
 [1] 10.220697  6.813117  4.541623  4.257346  3.405536
2.837663  2.918447  2.553385  2.269449  2.042299
```

THE TABLE

```
> data.frame(p0,P,c0,n,a1,t1)
      p0          P       c0 n        a1        t1
1  1e-04 0.13157895 9999.000 3 4088.2788 10.220697
2  2e-04 0.06172840 4999.000 4 2725.2466  6.813117
3  3e-04 0.03875969 3332.333 4 1816.6494  4.541623
4  4e-04 0.02906977 2499.000 5 1702.9384  4.257346
5  5e-04 0.02272727 1999.000 5 1362.2144  3.405536
6  6e-04 0.01865672 1665.667 5 1135.0651  2.837663
7  7e-04 0.01607717 1427.571 6 1167.3787  2.918447
8  8e-04 0.01392758 1249.000 6 1021.3542  2.553385
9  9e-04 0.01228501 1110.111 6  907.7795  2.269449
10 1e-03 0.01098901  999.000 6  816.9198  2.042299
```

-A.2-

```
######### SIMULATION FOR r=5 ###########

## The new number of false Alarms ##
> r<-5

#Calculating the proper number of n from (5.10) and
(5.14)#
> P1<-function(n){
+ r/(b*p0[1]-n+1)}
> n(P1(1))
[1] 0.8948694  # Thus n=1 #
----------------------------

> P2<-function(n){
+ r/(b*p0[2]-n+1)}
> n(P2(1))
[1] 1.845739
> n(P2(2))
[1] 1.772362  # Thus n=2 #
----------------------------

> P3<-function(n){
+ r/(b*p0[3]-n+1)}
> n(P3(2))
[1] 2.353483
> n(P3(3))
[1] 2.303228
> n(P3(2))
[1] 2.353483  # Thus n=2 #
-----------------------------

> P4<-function(n){
+ r/(b*p0[4]-n+1)}
> n(P4(2))
[1] 2.760411
> n(P4(3))
[1] 2.723232  # Thus n=3 #
-------------------------------

> P5<-function(n){
+  r/(b*p0[5]-n+1)}
> n(P5(5))
[1] 2.983357
> n(P5(3))
[1] 3.044336  # Thus n=3 #
-------------------------------

> P6<-function(n){
```

```
+ r/(b*p0[6]-n+1)}
> n(P6(3))
[1] 3.304352  # Thus n=3 #
--------------------------------

> P7<-function(n){
+ r/(b*p0[7]-n+1)}
> n(P7(3))
[1] 3.52285
> n(P7(4))
[1] 3.501647   # Thus n=4 #
---------------------------------

> P8<-function(n){
+ r/(b*p0[8]-n+1)}
> n(P8(3))
[1] 3.71128
> n(P8(4))
[1] 3.692817  # Thus n=4 #
--------------------------------

> P9<-function(n){
+ r/(b*p0[9]-n+1)}
> n(P9(4))
[1] 3.860576  # Thus n=4 #
-------------------------------

> P10<-function(n){
+ r/(b*p0[10]-n+1)}
> n(P10(4))
[1] 4.010036  # Thus n=4 #
-------------------------------

#####The table with the proper n and P #####
> a<-matrix(c(0),10,2)
> a[1,1]<-1
> a[1,2]<-P1(3)
> a[2,1]<-2
> a[2,2]<-P2(2)
> a[3,1]<-2
> a[3,2]<-P3(2)
> a[4,1]<-3
> a[4,2]<-P4(3)
> a[5,1]<-3
> a[5,2]<-P5(3)
> a[6,1]<-3
> a[6,2]<-P6(3)
> a[7,1]<-4
> a[7,2]<-P7(4)
> a[8,1]<-4
> a[8,2]<-P8(4)
```

```
> a[9,1]<-4
> a[9,2]<-P9(4)
> a[10,1]<-4
> a[10,2]<-P10(4)
> a
      [,1]       [,2]
 [1,]    1 0.65789474
 [2,]    2 0.27472527
 [3,]    2 0.17985612
 [4,]    3 0.13736264
 [5,]    3 0.10869565
 [6,]    3 0.08992806
 [7,]    4 0.07788162
 [8,]    4 0.06775068
 [9,]    4 0.05995204
[10,]    4 0.05376344
```

The n

```
> n_r5<-c(a[,1])
```

The P

```
> P_r5<-c(a[,2])
```

####Calculating the number of the expected births in a sequence which signals an alarm after the increase from (5.15)####

```
> e_r5<-c0*n_r5/7
> q<-1-(4.62/(exp(4.62)-1))
> a1_r5<-e_r5*q
```

####Calculating the expected time duration in which the a1 infants are born####

```
> t1_r5<-a1_r5/400
```

THE TABLE

```
> data.frame(p0,P_r5,c0,n_r5,a1_r5,t1_r5)
      p0       P_r5       c0 n_r5     a1_r5    t1_r5
1  1e-04 0.65789474 9999.000    1 1362.7596 3.406899
2  2e-04 0.27472527 4999.000    2 1362.6233 3.406558
3  3e-04 0.17985612 3332.333    2  908.3247 2.270812
4  4e-04 0.13736264 2499.000    3 1021.7630 2.554408
5  5e-04 0.10869565 1999.000    3  817.3287 2.043322
6  6e-04 0.08992806 1665.667    3  681.0391 1.702598
7  7e-04 0.07788162 1427.571    4  778.2525 1.945631
8  8e-04 0.06775068 1249.000    4  680.9028 1.702257
9  9e-04 0.05995204 1110.111    4  605.1863 1.512966
10 1e-03 0.05376344  999.000    4  544.6132 1.361533
```

-A.3-

```
#### Second Way to find n ####

## Comparison function from equation (5.7) ##
> r<-1
> Pcomp<-function(n){
+ (1-exp(-k))^n
+ }

#### Tables of comparison for each normal rate of
disease ####
> z<-matrix(c(0),6,3)
> z[,1]<-1:6
> for (i in 1:6){
+ z[i,2]<-P1(i)
+ z[i,3]<-Pcomp(i)
+ }
> z
     [,1]      [,2]       [,3]
[1,]    1 0.1041667 0.48211393
[2,]    2 0.1162791 0.23243384
[3,]    3 0.1315789 0.11205959
[4,]    4 0.1515152 0.05402549
[5,]    5 0.1785714 0.02604644
[6,]    6 0.2173913 0.01255735
### Smallest difference for n=3 ###

------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P2(i)
+ }
> z
     [,1]       [,2]       [,3]
[1,]    1 0.05208333 0.48211393
[2,]    2 0.05494505 0.23243384
[3,]    3 0.05813953 0.11205959
[4,]    4 0.06172840 0.05402549
[5,]    5 0.06578947 0.02604644
[6,]    6 0.07042254 0.01255735
### Smallest difference for n=4 ###

------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P3(i)
+ }
> z
```

```
     [,1]       [,2]       [,3]
[1,]    1 0.03472222 0.48211393
[2,]    2 0.03597122 0.23243384
[3,]    3 0.03731343 0.11205959
[4,]    4 0.03875969 0.05402549
[5,]    5 0.04032258 0.02604644
[6,]    6 0.04201681 0.01255735
### Smallest difference for n=5 ###

------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P4(i)
+ }
> z
     [,1]       [,2]       [,3]
[1,]    1 0.02604167 0.48211393
[2,]    2 0.02673797 0.23243384
[3,]    3 0.02747253 0.11205959
[4,]    4 0.02824859 0.05402549
[5,]    5 0.02906977 0.02604644
[6,]    6 0.02994012 0.01255735
### Smallest difference for n=5 ###

------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P5(i)
+ }
> z
     [,1]       [,2]       [,3]
[1,]    1 0.02083333 0.48211393
[2,]    2 0.02127660 0.23243384
[3,]    3 0.02173913 0.11205959
[4,]    4 0.02222222 0.05402549
[5,]    5 0.02272727 0.02604644
[6,]    6 0.02325581 0.01255735
### Smallest difference for n=5 ###

------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P6(i)
+ }
> z
     [,1]       [,2]       [,3]
[1,]    1 0.01736111 0.48211393
[2,]    2 0.01766784 0.23243384
[3,]    3 0.01798561 0.11205959
[4,]    4 0.01831502 0.05402549
[5,]    5 0.01865672 0.02604644
```

```
[6,]    6 0.01901141 0.01255735
### Smallest difference for n=6###

------------------------------------------------------------
--

> for (i in 1:6){
+ z[i,2]<-P7(i)
+ }
> z
     [,1]       [,2]       [,3]
[1,]    1 0.01488095 0.48211393
[2,]    2 0.01510574 0.23243384
[3,]    3 0.01533742 0.11205959
[4,]    4 0.01557632 0.05402549
[5,]    5 0.01582278 0.02604644
[6,]    6 0.01607717 0.01255735
### Smallest difference for n=6 ###

------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P8(i)
+ z[i,3]<-Pcomp(i)
+ }
> z
     [,1]       [,2]        [,3]
[1,]    1 0.01302083 0.482113928
[2,]    2 0.01319261 0.232433840
[3,]    3 0.01336898 0.112059592
[4,]    4 0.01355014 0.054025490
[5,]    5 0.01373626 0.026046441
[6,]    6 0.01392758 0.012557352
### Smallest difference for n=6 ###

------------------------------------------------------------
--

> for (i in 1:6){
+ z[i,2]<-P9(i)
+ }
> z
     [,1]       [,2]        [,3]
[1,]    1 0.01157407 0.482113928
[2,]    2 0.01170960 0.232433840
[3,]    3 0.01184834 0.112059592
[4,]    4 0.01199041 0.054025490
[5,]    5 0.01213592 0.026046441
[6,]    6 0.01228501 0.012557352
### Smallest difference for n=6 ###
```

```
------------------------------------------------------------
--
> for (i in 1:6){
+ z[i,2]<-P10(i)
+ }
> z
     [,1]       [,2]        [,3]
[1,]    1 0.01041667 0.482113928
[2,]    2 0.01052632 0.232433840
[3,]    3 0.01063830 0.112059592
[4,]    4 0.01075269 0.054025490
[5,]    5 0.01086957 0.026046441
[6,]    6 0.01098901 0.012557352
### Smallest difference for n=6 ###

------------------------------------------------------------
--
## The new n ##
> n<-c(3,4,5,5,5,6,6,6,6,6)

## The new a1 ##
> e<-c0*n/7
> q<-1-(4.62/(exp(4.62)-1))
> a1<-e*q
```

-A.4-

```
####### A third way to find n #######

### The function of M ###
> M<-function(k,n){
+ (1-(1-exp(-k))^n)/(exp(-k)*(1-exp(-k))^n)
+ }

### The function of gamma ###
> gamma<-function(k,n){
+ (-log(1-0.95^(1/n)))/k
+ }


### Find k for n=1,2,...,8 and M=48 ###
> f<-seq(0.001,2,by=0.001)
>kap<-c(0)
>m<-c(0)
>g<-c(0)
> for (j in 1:8){
+ q<-M(f,j)-48
+ for (i in 1:2000){
+ if (q[i]<0){
+ q[i]<-q[i]*(-1)
+ }
+ }
+ for (i in 1:2000){
+ if (q[i]==min(q)){
+  kappa<-i*0.001
+ }
+ }
+ kap[j]<-kappa  # The kappas #
+ m[j]<-M(kap[j],j)  # The M's #
+ g[j]<-gamma(kap[j],j) # The gammas respectively #
+ }

### The table with the M,k and gammas ###
> data.frame(m,kap,g)

### The table with the M,k and gammas ###
> data.frame(m,kap,g)
         m   kap          g
1 48.12080 0.021 142.653918
2 47.79204 0.169  21.752298
3 48.06186 0.368  11.079740
4 48.06683 0.569   7.667653
5 48.05415 0.759   6.040525
6 48.02193 0.936   5.092122
7 47.94634 1.101   4.468453
```

```
8 48.00808 1.254   4.029379

### The appropriate value of n ###
> n<-matrix(10,8,1)
> d<-c(0)
> for (i in 1:8){
+ d[i]<-g[i]-g[i+1]
+ if (d[i]<1){
+ n[i,1]<-i
+ }
+ }
> n<-min(n)
> n
[1] 5 # Thus n=5 #
```

-A.5-

```
##### The Sets Method Applied for a Large Scale Area#####

## The baseline rate values ##
> d<-c(1:10)
> p0<-d/10000

## The Expected Size of a set ##
> c0<-(1-p0)/p0
> c0

## The function of kappa ##
> k<-function(q,m){
+ -log(1-(q^(1/m)))
+ }


## The Probability of a False Alarm ##
> y<-40
> d<-4
> P0<-1/(40*4)

## The Probability of a False Alarm given that at least one set is completed ##
> N<-12000
> q0<-P0/(1-(1-p0)^N)
> q0

## kappa for m=5 completed sets ##
> kappa<-k(q0,5)
> kappa

## The Limits of the sets which define the in-control and out-of-control state ##
> kappa*c0
```

Appendix B

The code (we used in S-plus and Minitab) for the simulation of the Cuscore method is given below:

```
######## Optimal Cuscore ########
### Find E1 for n=1 ###
## s-plus ##
> c<-c(1:10)
> lamda<-c/10000
> D<-400*lamda*240
> D
 [1]  9.6 19.2 28.8 38.4 48.0 57.6 67.2 76.8 86.4 96.0

## Minitab ##
#c1=λο # c2= D(from s-plus)#
MTB > let c3=1/c2 #p0_n=1
MTB > let c4=-ln(1-c3) #kappa
MTB > let c5=1-exp(-7*c4) #p1
MTB > let c6=1/c5 #E1_n=1
------------------------------------------------------
--
### Find E1 for n=2 ###
## Minitab ##
MTB > let c8=(1+c2-(c2*c3))^(-1/2)######c8
MTB > let c9=(1+c2-(c2*c8))^(-1/2)######c9
MTB > let c10=(1+c2-(c2*c9))^(-1/2)######c10
MTB > let c11=(1+c2-(c2*c10))^(-1/2)######p0_n=2
MTB > let c12=-ln(1-c11)###kappa
MTB > let c13=1-exp(-7*c12)####p1

## S-plus ##
> p1<-
c(0.964011,0.873374,0.798358,0.739668,0.692703,0.654115
,0.621687,0.593927,0.569801,0.548569)
> E1<-function(p,n){
+ (n/((2*p)-1))-(((1-p)/(((2*p)-1)^2))*(1-((1-p)/p)^n))
+ }
> E1(p1,2)
 [1] 2.113391 2.455975 2.821505 3.179748 3.527659
3.865961 4.195884 4.518583
 [9] 4.835019 5.145979



## Minitab ##
in the column c14 we place the E1 from s-plus
------------------------------------------------------
--
```

```
### Find E1 for n=3 ###
## Minitab ##
MTB > let c16=(1+c2-(c2*c11))^(-1/3)######c8
MTB > let c17=(1+c2-(c2*c16))^(-1/3)######c9
MTB > let c18=(1+c2-(c2*c17))^(-1/3)######c10
MTB > let c19=(1+c2-(c2*c18))^(-1/3)######p0_n=2
MTB > let c20=-ln(1-c19)###kappa
MTB > let c21=1-exp(-7*c20)####p1

## S-plus ##
> p1<-
c(0.997797,0.982580,0.962546,0.942624,0.923994,0.906847
,0.891104,0.876621,0.863248,0.850852)
> length(p1)
[1] 10
> E1(p1,3)
 [1] 3.011054 3.089593 3.199158 3.315682 3.432146
3.546348 3.657643 3.765934
 [9] 3.871340 3.974031

## Minitab ##
in the column c22 we place the E1 from s-plus
----------------------------------------------------
--

### Find E1 for n=4 ###
## Minitab ##
MTB > let c24=(1+c2-(c2*c19))^(-1/4)######c8
MTB > let c25=(1+c2-(c2*c24))^(-1/4)######c9
MTB > let c26=(1+c2-(c2*c25))^(-1/4)######c10
MTB > let c27=(1+c2-(c2*c26))^(-1/4)######p0_n=2
MTB > let c28=-ln(1-c27)###kappa
MTB > let c29=1-exp(-7*c28)####p1

## S-plus ##
> p1<-
c(0.999858,0.997447,0.992814,0.987247,0.981382,0.975504
,0.969742,0.964155,0.958764,0.953575)
> E1(p1,4)
 [1] 4.000994 4.017950 4.050929 4.091265 4.134619
4.178979 4.223375 4.267310
 [9] 4.310558 4.353000

## Minitab ##
in the column c30 we place the E1 from s-plus
```

Appendix C

The code (we used in S-plus) in order to find the out-of-control expected delay of the sets method is given below:

```
#### Comparison Sets vs Cuscore ####

> D1<-function(p,n){
+ (1-(p^n))/((p^n)*(1-p))
+ }
> p1<-
c(0.536990,0.873374,0.798358,0.739668,0.923994,0.906847
,0.891104,0.876621,0.863248,0.850852)
> n1<-c(1,2,2,2,3,3,3,3,3,3)

# The out-of-control Expected Delay for the Sets method
#
> a<-c(0)
> for (i in 1:10){
+ a[i]<-D1(p1[i],n1[i])
+ }
> a
 [1] 1.862232 2.455975 2.821505 3.179748 3.521171
3.659622 3.794781 3.926486
 [9] 4.054852 4.180051
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Appendix D

The tables below are given in Lucas(1985) and are calculated by using the Markov chain approach discussed by Brook and Evans(1972).

-D.1-

Poisson CUSUM's Average Run Lengths
(Increasing Rate Case, With the FIR feature)

			Mean as a Multiple of k													
h	*k*	S_0	*.1*	*.2*	*.3*	*.4*	*.5*	*.6*	*.8*	*1.0*	*1.2*	*1.4*	*1.7*	*2.0*	*2.5*	*3.0*	*5.0*
1.0	.25	.5	494.	133.	63.5	38.2	26.0	19.2	12.1	8.65	6.65	5.37	4.17	3.41	2.65	2.20	1.46
2.0	.25	1.0	10,805.	1,410.	443.	200.	110.	68.5	34.0	16.9	14.2	10.6	7.59	5.90	4.35	3.49	2.14
3.0	.25	2.0	*	17,505.	3,220.	975.	392.	190.	65.3	31.2	18.4	12.5	8.26	6.17	4.42	3.52	2.14
5.0	.25	3.0	*	*	*	27,205.	5,670.	1,630.	263.	80.9	37.8	22.8	13.9	9.98	8.12	5.32	3.00
7.0	.25	4.0	*	*	*	*	72,905.	11,805.	813.	155.	59.3	33.3	19.6	14.0	11.3	7.26	3.92
10.0	.25	5.0	*	*	*	*	*	*	3,670.	333.	101.	53.6	31.1	22.0	17.7	11.3	5.89
1.0	.50	.5	283.	75.2	35.5	21.2	14.4	10.6	6.67	4.47	3.70	3.02	2.40	2.01	1.80	1.42	1.11
2.0	.50	1.0	6,874.	872.	266.	117.	62.7	38.5	18.6	11.1	7.58	5.66	4.07	3.20	2.42	1.99	1.34
3.0	.50	2.0	*	11,305.	2,000.	582.	227.	107.	35.3	16.5	9.64	6.55	4.37	3.32	2.44	2.01	1.34
5.0	.50	3.0	*	*	*	16,305.	3,280.	910.	140.	41.8	19.3	11.7	7.16	5.20	4.27	2.88	1.73
7.0	.50	4.0	*	*	*	*	42,205.	6,570.	430.	79.4	30.0	16.9	10.0	7.18	5.84	3.84	2.18
10.0	.50	5.0	*	*	*	*	*	*	1,940.	169.	50.7	27.0	15.7	11.2	7.62	5.84	3.16
15.0	.50	8.0	*	*	*	*	*	*	18,205.	350.	72.7	37.2	21.5	15.2	10.3	7.84	4.16
20.0	.50	10.0	*	*	*	*	*	*	*	638.	103.	52.2	30.0	21.2	14.3	10.8	5.67
2.0	1.0	1.0	5,620.	670.	193.	80.4	41.3	24.3	11.1	6.36	4.25	3.15	2.29	1.84	1.46	1.27	1.04
3.0	1.0	2.0	*	8,970.	1,480.	404.	149.	66.9	20.4	9.06	5.19	3.52	2.40	1.88	1.64	1.27	1.04
5.0	1.0	3.0	*	*	*	11,300.	2,140.	562.	78.9	22.2	10.0	6.02	3.74	2.76	2.00	1.63	1.14
7.0	1.0	4.0	*	*	*	*	27,500.	4,050.	240.	41.5	15.3	8.61	5.16	3.74	2.64	2.09	1.30
10.0	1.0	5.0	*	*	*	*	*	72,200.	1,080.	87.5	25.6	13.7	8.02	5.74	3.98	3.09	1.76
15.0	1.0	8.0	*	*	*	*	*	*	10,100.	179.	36.6	18.7	10.9	7.74	5.31	4.09	2.28
20.0	1.0	10.0	*	*	*	*	*	*	89,900.	324.	51.8	26.2	15.2	10.7	7.31	5.59	3.03
2.0	2.0	1.0	17,200.	1,200.	259.	89.3	40.0	21.2	8.40	4.49	2.92	2.16	1.61	1.35	1.15	1.07	1.00
3.0	2.0	2.0	*	13,600.	1,780.	418.	136.	55.5	14.7	6.07	3.41	2.34	1.66	1.36	1.15	1.07	1.00
5.0	2.0	3.0	*	*	*	11,000.	1,830.	432.	51.2	13.2	5.78	3.50	2.25	1.73	1.34	1.17	1.01
7.0	2.0	4.0	*	*	*	*	23,300.	3,060.	150.	23.4	8.42	4.76	2.92	2.18	1.61	1.33	1.03
10.0	2.0	5.0	*	*	*	*	*	54,100.	657.	47.5	13.6	7.29	4.35	3.17	2.25	1.79	1.13
15.0	2.0	8.0	*	*	*	*	*	*	6,100.	94.7	19.0	9.81	5.78	4.17	2.93	2.31	1.34
20.0	2.0	10.0	*	*	*	*	*	*	54,200.	169.	26.6	13.6	7.92	5.67	3.93	3.06	1.76
2.0	3.0	1.0	63,000.	2,470.	400.	115.	45.2	21.8	7.64	3.83	2.43	1.80	1.38	1.19	1.06	1.02	1.00
3.0	3.0	2.0	*	24,200.	2,490.	501.	146.	54.7	12.8	4.99	2.75	1.91	1.40	1.20	1.06	1.02	1.00
5.0	3.0	3.0	*	*	*	12,100.	1,850.	403.	41.8	9.99	4.31	2.63	1.75	1.39	1.15	1.06	1.00
7.0	3.0	4.0	*	*	*	*	23,400.	2,820.	120.	17.1	6.05	3.45	2.17	1.66	1.28	1.12	1.00
10.0	3.0	5.0	*	*	*	*	*	49,600.	514.	33.8	9.47	5.12	3.10	2.29	1.66	1.36	1.02
15.0	3.0	8.0	*	*	*	*	*	*	4,740.	66.0	13.1	6.80	4.06	2.96	2.12	1.69	1.07
20.0	3.0	10.0	*	*	*	*	*	*	42,000.	117.	18.2	9.30	5.49	3.97	2.79	2.21	1.27
2.0	5.0	1.0	*	12,000.	1,060.	211.	64.8	26.2	7.29	3.27	2.00	1.49	1.19	1.07	1.02	1.00	1.00
3.0	5.0	2.0	*	96,500.	5,670.	822.	193.	61.5	11.6	4.05	2.18	1.54	1.20	1.08	1.02	1.00	1.00
5.0	5.0	3.0	*	*	*	16,200.	2,140.	409.	34.4	7.30	3.08	1.93	1.36	1.15	1.04	1.01	1.00
7.0	5.0	4.0	*	*	*	*	26,400.	2,810.	94.9	11.9	4.08	2.37	1.56	1.26	1.07	1.02	1.00
10.0	5.0	5.0	*	*	*	*	*	49,200.	397.	22.4	6.11	3.35	2.08	1.58	1.21	1.07	1.00
15.0	5.0	8.0	*	*	*	*	*	*	3,630.	42.5	8.27	4.36	2.66	1.98	1.45	1.19	1.00
20.0	5.0	10.0	*	*	*	*	*	*	32,100.	73.8	11.3	5.86	3.52	2.59	1.87	1.49	1.01
2.0	7.0	1.0	*	61,300.	2,950.	403.	96.7	32.9	7.41	3.01	1.79	1.35	1.11	1.03	1.00	1.00	1.00
3.0	7.0	2.0	*	*	14,300.	1,450.	271.	73.9	11.4	3.61	1.90	1.37	1.11	1.03	1.00	1.00	1.00
5.0	7.0	3.0	*	*	*	23,900.	2,650.	450.	31.7	6.07	2.52	1.62	1.20	1.07	1.01	1.00	1.00
7.0	7.0	4.0	*	*	*	*	30,600.	2,990.	84.8	9.50	3.21	1.91	1.32	1.12	1.02	1.00	1.00
10.0	7.0	5.0	*	*	*	*	*	52,600.	349.	17.3	4.64	2.57	1.64	1.29	1.07	1.01	1.00
15.0	7.0	8.0	*	*	*	*	*	*	3,170.	32.2	6.17	3.29	2.04	1.54	1.17	1.04	1.00
20.0	7.0	10.0	*	*	*	*	*	*	27,900.	55.0	8.34	4.37	2.67	1.99	1.44	1.16	1.00
2.0	10.0	1.0	*	*	14,000.	1,090.	180.	47.4	7.89	2.80	1.61	1.23	1.05	1.01	1.00	1.00	1.00
3.0	10.0	2.0	*	*	61,600.	3,610.	473.	102.	11.8	3.26	1.68	1.24	1.05	1.01	1.00	1.00	1.00
5.0	10.0	3.0	*	*	*	48,500.	3,960.	554.	30.3	5.10	2.08	1.38	1.09	1.02	1.00	1.00	1.00
7.0	10.0	4.0	*	*	*	*	90,100.	3,440.	78.4	7.63	2.54	1.56	1.15	1.04	1.00	1.00	1.00
10.0	10.0	5.0	*	*	*	*	*	59,200.	316.	13.3	3.50	1.99	1.33	1.11	1.01	1.00	1.00
15.0	10.0	8.0	*	*	*	*	*	*	2,850.	24.2	4.56	2.47	1.58	1.23	1.04	1.00	1.00
20.0	10.0	10.0	*	*	*	*	*	*	25,100.	40.6	6.08	3.23	2.01	1.52	1.13	1.02	1.00

*Value greater than 10^5.

-D.2-

Poisson CUSUM's Average Run Lengths
(Increasing Rate Case, Without the FIR feature)

			Mean as a Multiple of k														
h	k	S_0	.1	.2	.3	.4	.5	.6	.8	1.0	1.2	1.4	1.7	2.0	2.5	3.0	5.0
1.0	.25	.0	518.	145.	71.5	44.3	31.0	23.4	15.4	11.3	8.94	7.38	5.88	4.91	3.89	3.26	2.11
2.0	.25	.0	10,800.	1,440.	460.	212.	120.	76.7	40.0	25.4	18.1	13.9	10.3	8.14	6.09	4.92	2.95
3.0	.25	.0	*	17,900.	3,390.	1,070.	454.	234.	91.0	48.5	31.1	22.4	15.6	11.9	8.61	6.79	3.84
5.0	.25	.0	*	*	*	27,800.	5,950.	1,790.	332.	120.	63.5	41.5	26.9	19.9	16.4	10.8	5.78
7.0	.25	.0	*	*	*	*	74,005.	12,205.	958.	223.	99.9	61.3	38.4	27.9	22.8	14.8	7.80
10.0	.25	.0	*	*	*	*	*	*	3,950.	438.	158.	91.2	55.5	39.9	32.4	20.8	10.8
1.0	.50	.0	297.	82.1	40.1	24.7	17.2	12.9	8.46	6.22	4.91	4.07	3.27	2.75	2.45	1.90	1.34
2.0	.50	.0	6,900.	888.	276.	124.	68.8	43.4	22.2	13.9	9.84	7.54	5.57	4.43	3.35	2.73	1.72
3.0	.50	.0	*	11,600.	2,110.	641.	264.	133.	49.8	26.1	16.6	11.9	8.29	6.36	4.63	3.68	2.16
5.0	.50	.0	*	*	*	16,705.	3,440.	1,000.	178.	62.7	33.0	21.5	14.0	10.4	7.29	5.68	3.14
7.0	.50	.0	*	*	*	*	42,805.	6,840.	509.	115.	51.2	31.4	19.7	14.4	9.96	7.68	4.14
10.0	.50	.0	*	*	*	*	*	*	2,080.	122.	80.2	46.4	28.3	20.4	14.0	10.7	5.64
15.0	.50	.0	*	*	*	*	*	*	18,905.	214.	130.	71.4	42.5	30.4	20.6	15.7	8.14
20.0	.50	.0	*	*	*	*	*	*	18,205.	350.	180.	96.4	58.2	40.4	27.3	20.7	10.6
2.0	1.0	.0	5,650.	684.	201.	86.3	45.8	27.9	13.6	8.21	5.70	4.32	3.17	2.52	1.93	1.60	1.13
3.0	1.0	.0	*	9,180.	1,560.	446.	174.	83.8	29.5	14.8	9.24	6.56	4.54	3.49	2.56	2.06	1.29
5.0	1.0	.0	*	*	96,800.	11,500.	2,250.	620.	102.	34.2	17.6	11.4	7.39	5.49	3.89	3.06	1.76
7.0	1.0	.0	*	*	*	*	27,900.	4,220.	286.	61.5	26.8	16.4	10.2	7.49	5.23	4.06	2.28
10.0	1.0	.0	*	*	*	*	*	72,700.	1,160.	117.	41.3	23.8	14.5	10.5	7.23	5.56	3.03
15.0	1.0	.0	*	*	*	*	*	*	10,500.	251.	66.2	36.3	21.7	15.5	10.6	8.06	4.28
20.0	1.0	.0	*	*	*	*	*	*	90,900.	434.	91.1	48.8	28.8	20.5	13.9	10.6	5.53
2.0	2.0	.0	17,200.	1,200.	264.	93.3	43.0	23.6	9.96	5.60	3.74	2.79	2.05	1.66	1.33	1.17	1.01
3.0	2.0	.0	*	13,700.	1,830.	444.	152.	65.6	19.9	9.27	5.57	3.90	2.70	2.10	1.59	1.33	1.03
5.0	2.0	.0	*	*	*	11,100.	1,900.	468.	64.3	19.8	9.87	6.34	4.12	3.09	2.23	1.79	1.13
7.0	2.0	.0	*	*	*	*	23,600.	3,160.	176.	34.3	14.5	8.82	55.5	4.09	2.91	2.30	1.34
10.0	2.0	.0	*	*	*	*	*	54,400.	705.	63.6	21.8	12.6	7.69	5.59	3.91	3.05	1.76
15.0	2.0	.0	*	*	*	*	*	*	6,320.	132.	34.3	18.8	11.3	8.09	5.57	4.30	2.39
20.0	2.0	.0	*	*	*	*	*	*	54,800.	226.	46.7	25.1	14.8	10.6	7.24	5.55	3.02
2.0	3.0	.0	63,100.	2,480.	404.	118.	47.7	23.7	8.85	4.66	3.02	2.23	1.65	1.37	1.15	1.06	1.00
3.0	3.0	.0	*	24,300.	2,520.	520.	158.	62.3	16.7	7.28	4.25	2.96	2.06	1.62	1.28	1.12	1.00
5.0	3.0	.0	*	*	*	12,200.	1,910.	430.	51.5	14.7	7.16	4.57	2.99	2.26	1.66	1.35	1.02
7.0	3.0	.0	*	*	*	*	23,700.	2,900.	139.	24.9	10.3	6.23	3.94	2.93	2.11	1.69	1.07
10.0	3.0	.0	*	*	*	*	*	49,800.	548.	45.1	15.2	8.73	5.37	3.93	2.79	2.21	1.27
15.0	3.0	.0	*	*	*	*	*	*	4,890.	92.2	23.5	12.9	7.75	5.60	3.89	3.04	1.78
20.0	3.0	.0	*	*	*	*	*	*	42,400.	156.	31.8	17.1	10.1	7.26	5.01	3.87	2.18
2.0	5.0	.0	*	12,000.	1,070.	214.	66.7	27.6	8.19	3.85	2.38	1.75	1.32	1.14	1.04	1.01	1.00
3.0	5.0	.0	*	96,500.	5,690.	836.	201.	67.0	14.3	5.57	3.12	2.15	1.53	1.25	1.07	1.02	1.00
5.0	5.0	.0	*	*	*	16,300.	2,180.	428.	41.1	10.4	4.87	3.09	2.04	1.57	1.21	1.07	1.00
7.0	5.0	.0	*	*	*	*	26,500.	2,870.	108.	17.0	6.78	4.09	2.61	1.97	1.44	1.19	1.00
10.0	5.0	.0	*	*	*	*	*	49,400.	422.	29.8	9.73	5.59	3.48	2.58	1.87	1.49	1.01
15.0	5.0	.0	*	*	*	*	*	*	3,740.	59.2	14.7	8.09	4.90	3.58	2.54	2.04	1.13
20.0	5.0	.0	*	*	*	*	*	*	32,300.	98.5	19.7	10.6	6.33	4.58	3.21	2.52	1.47
2.0	7.0	.0	*	61,300.	2,950.	406.	98.4	34.2	8.16	3.47	2.07	1.52	1.19	1.07	1.01	1.00	1.00
3.0	7.0	.0	*	*	14,300.	1,460.	278.	78.4	13.6	4.79	2.59	1.78	1.30	1.12	1.02	1.00	1.00
5.0	7.0	.0	*	*	*	24,000.	2,680.	465.	37.0	8.47	3.83	2.43	1.63	1.29	1.07	1.01	1.00
7.0	7.0	.0	*	*	*	*	30,800.	3,040.	95.5	13.4	5.20	3.14	2.02	1.54	1.17	1.04	1.00
10.0	7.0	.0	*	*	*	*	*	52,600.	369.	23.0	7.32	4.21	2.65	1.99	1.44	1.16	1.00
15.0	7.0	.0	*	*	*	*	*	*	3,250.	44.6	10.9	6.00	3.67	2.71	1.97	1.57	1.01
20.0	7.0	.0	*	*	*	*	*	*	28,100.	73.4	14.4	7.79	4.69	3.42	2.44	1.98	1.07
2.0	10.0	.0	*	*	14,000.	1,090.	181.	48.5	8.53	3.16	1.81	1.34	1.09	1.02	1.00	1.00	1.00
3.0	10.0	.0	*	*	61,600.	3,610.	479.	105.	13.5	4.16	2.16	1.50	1.15	1.04	1.00	1.00	1.00
5.0	10.0	.0	*	*	*	48,500.	3,980.	566.	34.5	6.90	3.01	1.91	1.32	1.11	1.01	1.00	1.00
7.0	10.0	.0	*	*	*	793,000.	40,200.	3,470.	86.8	10.6	3.97	2.40	1.57	1.23	1.04	1.00	1.00
10.0	10.0	.0	*	*	*	*	*	59,400.	332.	17.6	5.45	3.15	2.01	1.51	1.13	1.02	1.00
15.0	10.0	.0	*	*	*	*	*	*	2,920.	33.4	7.95	4.41	2.73	2.06	1.48	1.16	1.00
20.0	10.0	.0	*	*	*	*	*	*	25,200.	54.1	10.4	5.65	3.44	2.55	1.88	1.48	1.00

*Value greater than 10^5.

-D.3-

The code we used in S-plus for the simulation of the Poisson CUSUM scheme:

```
###### POISSON CUSUM ######

## Generating values from Poisson ##
> p1<-rpois(20,4)
> p1
 [1] 2 3 2 2 2 5 5 3 4 1 1 5 2 8 4 2 3 5 7 6
> p2<-rpois(20,7)
> p<-c(p1,p2)
> p
 [1]  2  3  2  2  2  5  5  3  4  1  1  5  2  8  4  2  3
5  7  6  6 10  8  6  6  4 10 10  7
[30] 14  2  9 12 15  9  6  4  5  6  2

## The CUSUM Scheme with and without the FIR ##
# With the FIR #
> POIS_CUS1<-function(k,r,h){
+ s<-c(0)
+ s[1]<-(h/2)+r[1]-k
+  if (s[1]<0) {s[1]<-0}
+ for (i in 2:40){
+ s[i]<-s[i-1]+r[i]-k
+ if (s[i]<0) {s[i]<-0}
+ }
+ plot(period,s,col=3,ylab="Cumulative Sum",xlab="Time
Period")
+ lines(period,s,col=3)
+ abline(h=(h),col=2)
+ }
> POIS_CUS1(5,p,10)
> abline(v=(23),lty=3,col=3)

# Without the FIR #
> POIS_CUS<-function(k,r){
+ s<-c(0)
+ for (i in 1:40){
+ s[i]<-s[i-1]+r[i]-k
+ if (s[i]<0) {s[i]<-0}
+ }
+ plot(period,s,col=3,ylab="Cumulative Sum",xlab="Time
Period")
+ lines(period,s,col=3)
+ abline(h=(10),col=2)
+ }
> POI_CUS(5,p)
> abline(v=(23),lty=3,col=3)
```


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Appendix E

-E.1-

❖ *Table for N(0,1) for each one of the nine areas.*

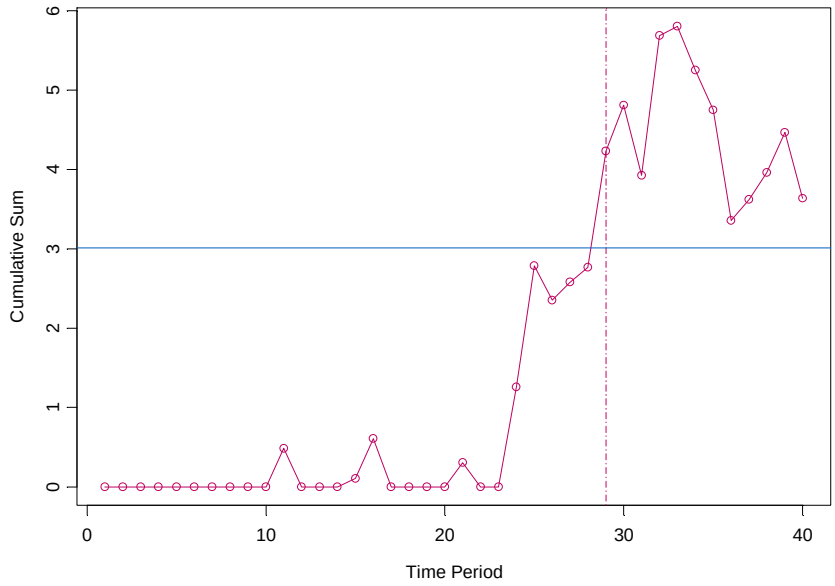
Period	*Region 1*	*Region 2*	*Region 3*	*Region 4*	*Region 5*	*Region 6*	*Region 7*	*Region 8*	*Region 9*
1	0,41582	-0,38144	0,71736	-0,31835	0,46841	-1,85771	-0,14439	0,66588	-1,45069
2	-1,34842	0,09742	-0,45174	-1,23156	-0,18584	0,78473	1,68392	-0,27911	-0,99563
3	-0,78504	0,68365	-0,65561	-0,20796	-0,22638	1,81656	1,48803	0,74188	1,45907
4	-0,83648	1,26462	1,02065	-0,15173	1,81522	-2,65010	-1,01649	0,13865	-1,47886
5	0,01403	-0,69039	0,81121	-1,32920	-1,01923	0,66423	-1,31545	-0,67074	-0,47185
6	-0,23819	0,10114	2,19297	0,67889	-0,77469	-0,61284	1,25016	-1,34498	-1,67495
7	0,45008	1,48601	-1,29020	1,20652	0,19245	1,10630	0,15131	0,51922	-0,23316
8	-1,82953	1,14686	-0,45899	-2,08894	-2,00393	0,40907	0,28793	0,43888	-1,23805
9	-0,20107	-0,35150	-1,02987	1,26068	0,15516	-0,23289	1,45883	1,00479	0,41722
10	-1,11774	1,25533	2,62136	0,72618	-0,28784	-0,92571	-0,07356	0,53368	1,42631
11	0,98532	0,66070	0,85998	-0,46840	-0,08254	1,13479	0,19762	-1,61812	0,21364
12	-0,88706	0,63594	0,85691	0,15308	-0,17487	0,68239	0,75395	-0,53415	1,73795
13	-0,74147	-0,13405	1,27770	-0,58923	1,97022	0,89292	-1,69801	-0,50158	-0,13859
14	0,49759	-1,57070	-0,24252	0,17270	0,45301	1,92679	-0,00760	0,39056	-0,79184
15	0,60502	1,56000	0,85329	0,58868	0,54826	-1,75153	1,28713	-0,24788	-0,29292
16	1,00473	-0,32744	-0,41565	1,24417	0,36710	1,39520	-3,53075	0,50563	-0,16031
17	-0,76336	0,50897	0,33033	-0,86251	0,17584	-0,88295	1,08108	1,36287	-1,19796
18	-2,53159	0,89491	0,06431	-0,56156	0,29516	-0,07153	0,67365	-1,81468	1,37303
19	-0,10022	-1,71465	-0,35631	-0,84478	0,01129	-0,62794	0,52465	-1,12899	0,67287
20	-0,29377	-0,01323	-0,67293	-0,32563	-1,08156	0,19716	-0,00327	-1,56588	-1,01454

❖ *Table after the change in the mean for each one of the nine areas.*

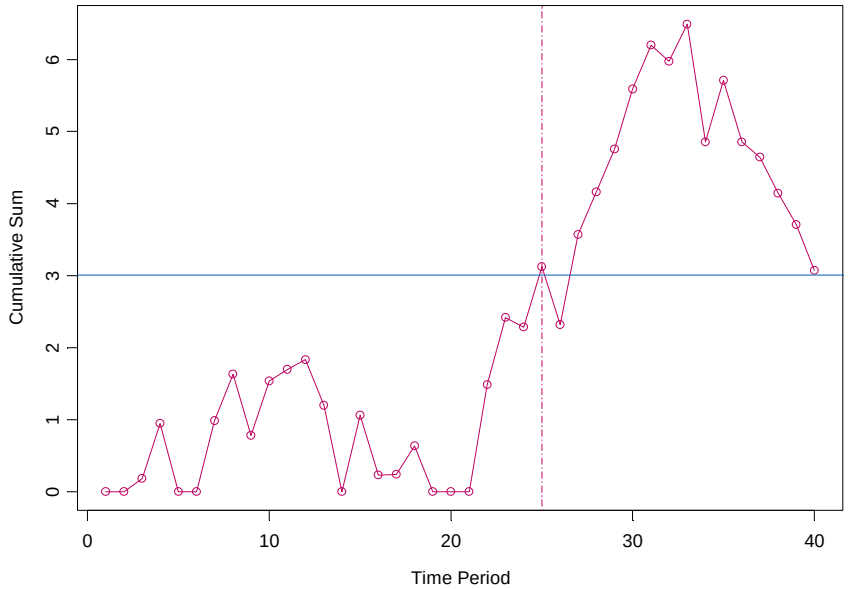
Period	*Region 1*	*Region 2*	*Region 3*	*Region 4*	*Region 5*	*Region 6*	*Region 7*	*Region 8*	*Region 9*
21	0,80559	-1,03404	0,03039	0,95801	2,26004	0,85243	1,05929	0,43797	2,97966
22	0,14105	1,98492	0,58935	-0,91913	0,89931	-1,51403	0,95604	0,25103	-0,37933
23	0,25698	1,43465	2,58853	-1,41943	2,00807	1,80059	1,10127	0,05865	-0,42117
24	1,75863	0,36656	-1,81370	0,54278	-0,66747	1,95028	-0,53798	3,12483	-0,40883
25	2,02860	1,33783	1,14479	-0,32970	0,53093	1,58163	0,29998	1,05387	1,34590
26	0,06401	-0,30641	-0,27185	0,91316	1,17525	0,20249	-0,45979	0,70425	1,00437
27	0,73032	1,75258	1,52275	0,70360	2,18047	-0,06315	1,00709	1,12978	-0,30273
28	0,68455	1,08982	0,42286	3,08482	1,11053	2,41280	-1,29459	-0,46265	-1,26922
29	1,96520	1,09619	-1,61709	-1,95850	1,77202	1,32315	-0,19347	0,99680	2,39611
30	1,07936	1,33340	0,48232	0,98567	-0,24031	1,92322	-1,25547	2,07574	1,78574
31	-0,38411	1,11066	-0,15097	1,69870	2,79923	-1,21268	-0,38346	0,03381	0,20704
32	2,26125	0,27410	-0,10126	1,65957	1,77736	2,71995	-1,73675	-1,98983	-0,53716
33	0,61521	1,01402	-0,65468	1,09882	-0,79483	-1,33701	0,49867	-0,33729	-0,94827
34	-0,05345	-1,13417	-0,69927	-0,18067	0,83489	0,12583	0,63097	-0,02851	-1,01407
35	-0,00009	1,35453	0,12516	-1,12081	1,26136	0,98994	-0,72824	0,97726	0,95611
36	-0,89270	-0,35450	-0,83167	1,23665	1,42278	1,12837	0,67222	0,09723	-0,35271
37	0,76269	0,29064	0,06562	0,06846	0,46300	0,76176	0,07138	0,49282	-0,54428
38	0,84122	-0,00086	0,26837	2,10915	0,97853	0,42393	-1,34082	1,11248	1,55765
39	1,00577	0,06092	-0,05394	1,66525	-0,35092	-0,24709	-1,31535	1,07484	0,51802
40	-0,32999	-0,13191	1,47785	1,99385	0,94120	-0,72778	1,55872	-0,43880	-0,93122

-E.2-

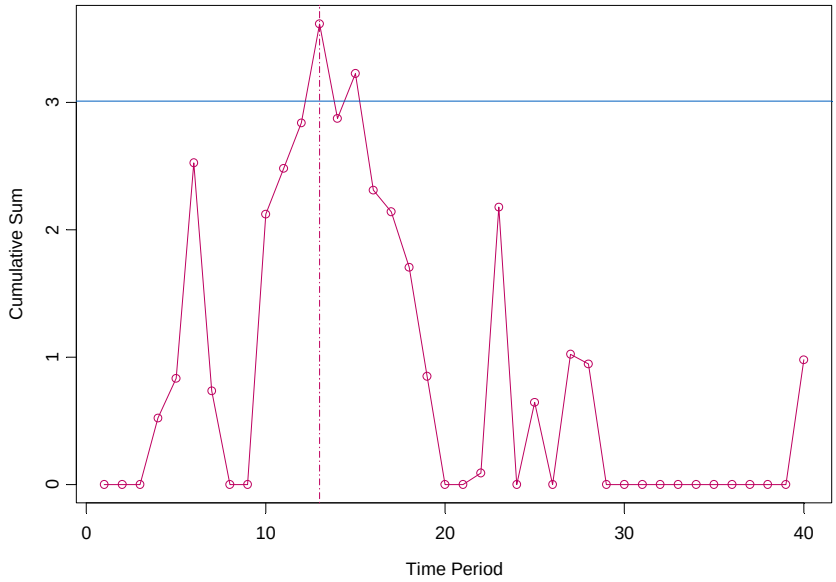
❖ *Region 1: Alarm at 29th Observation*



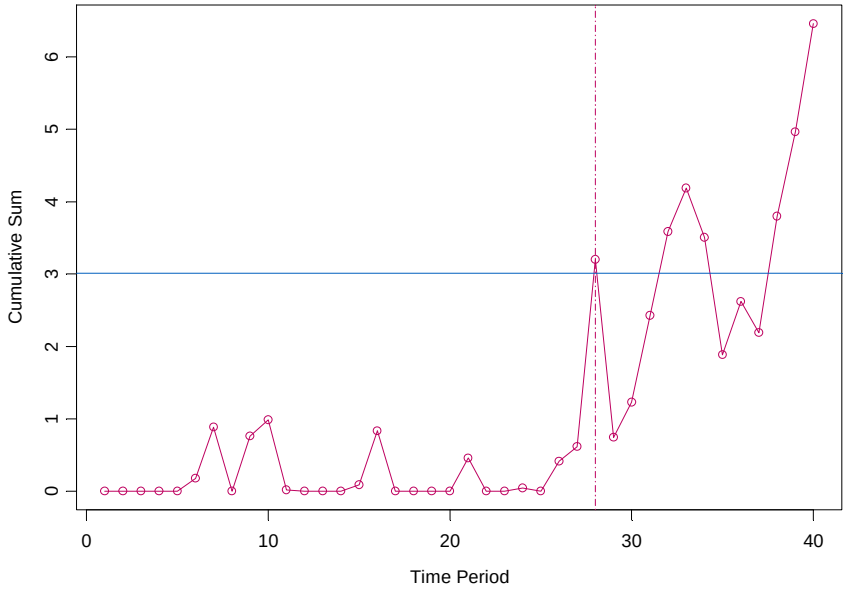
❖ *Region 2: Alarm at 25th Observation*



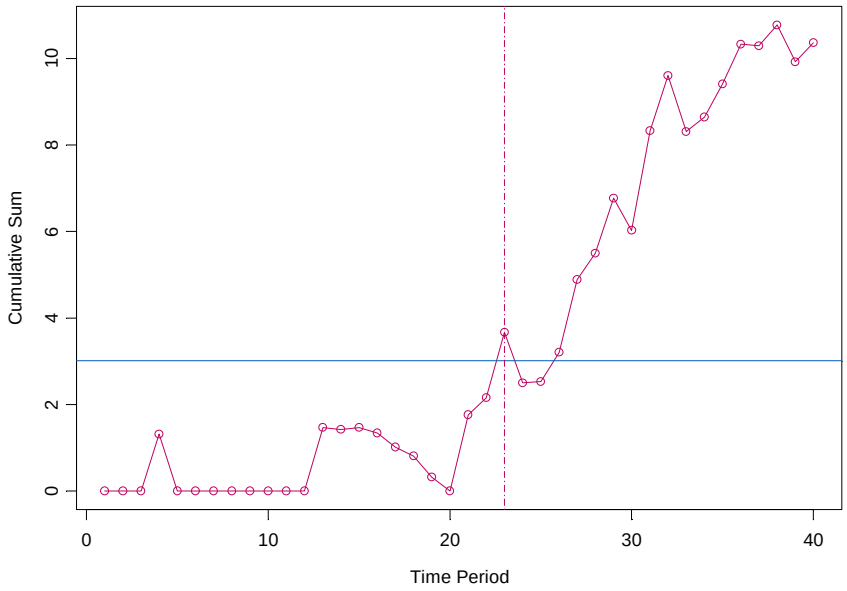
❖ *Region 3: False Alarm at 13th Observation*



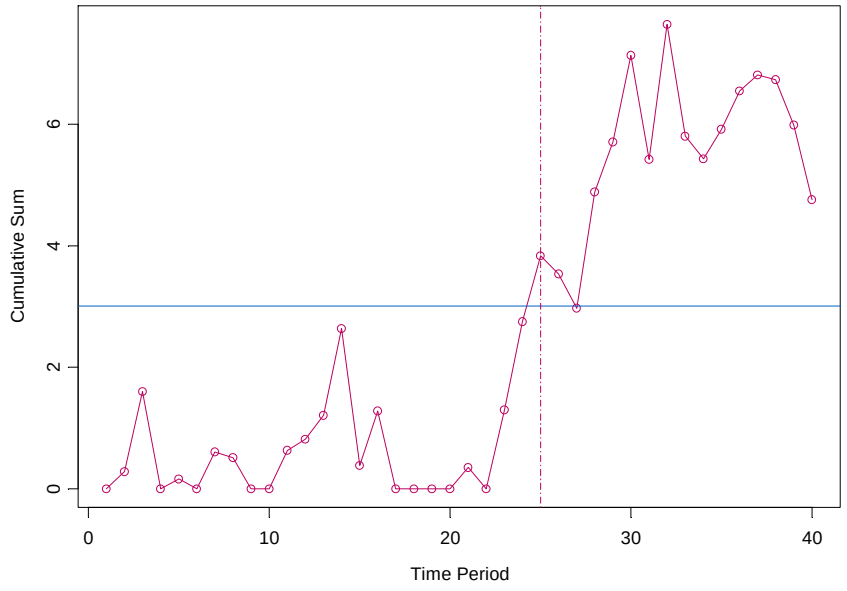
❖ *Region 4: Alarm at 28th Observation*



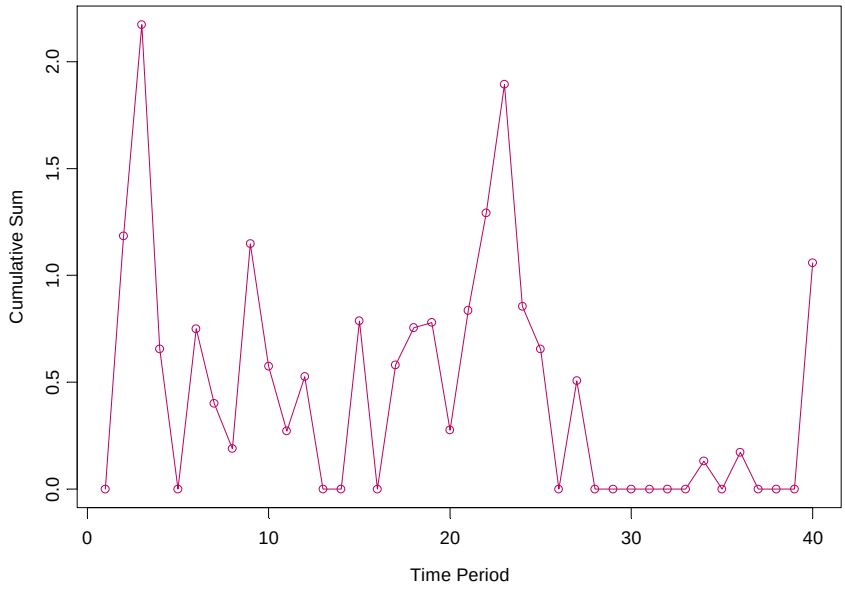
❖ *Region 5: Alarm at 23rd Observation*



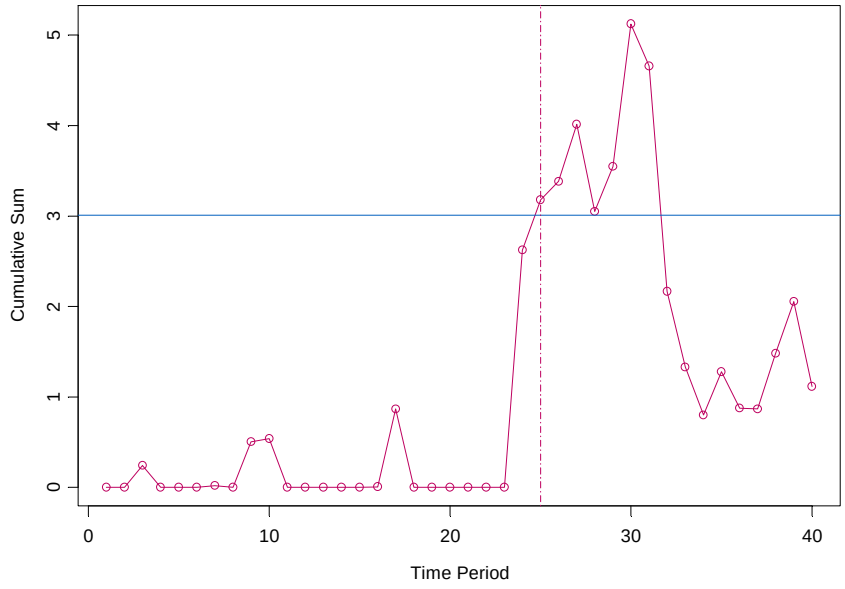
❖ *Region 6: Alarm at 25th Observation*



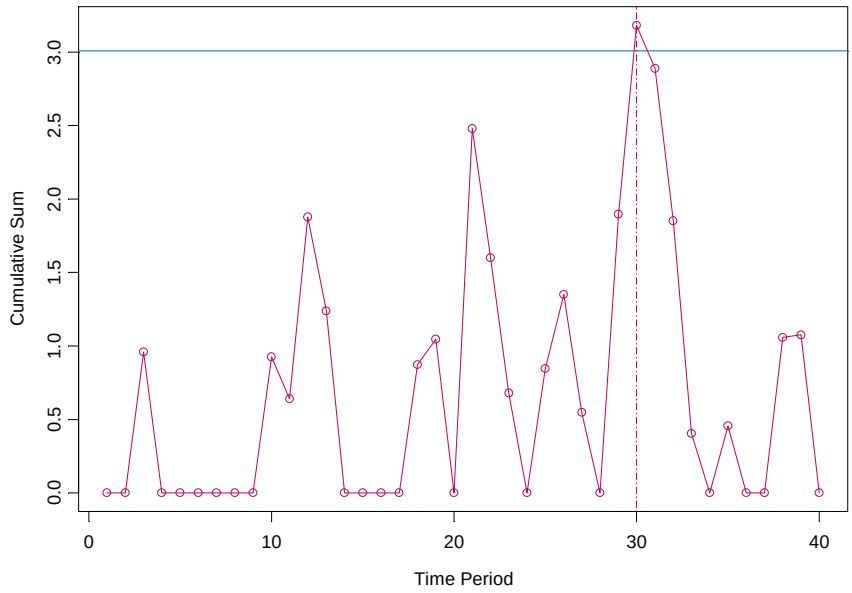
❖ *Region 7: No Alarm*



❖ *Region 8: Alarm at 25th Observation*



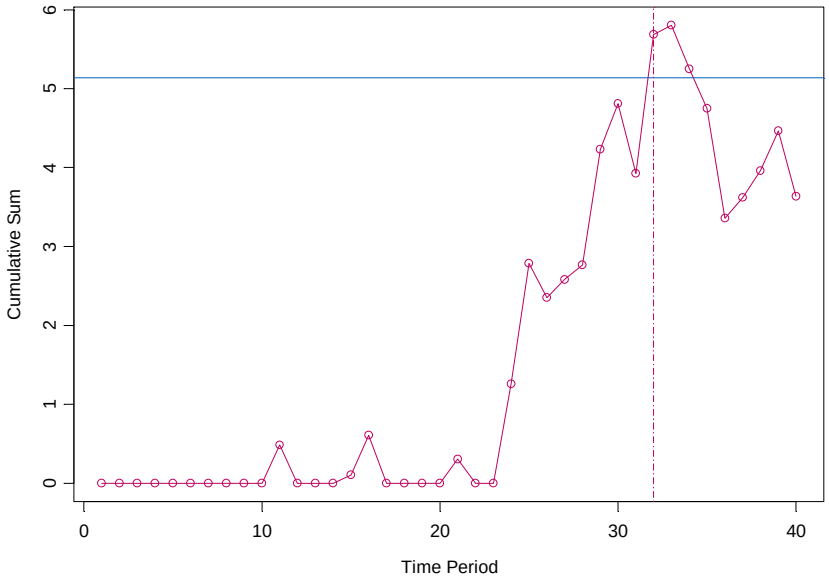
❖ *Region 9: Alarm at 30th Observation*



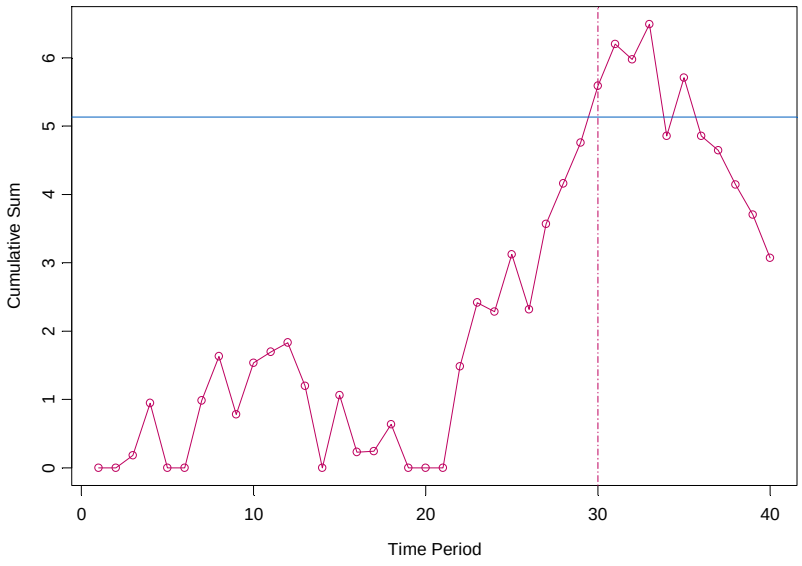
-E.3-

The CUSUM schemes for each region for the new ARL=1075.99 and the new h=5.135 are given below:

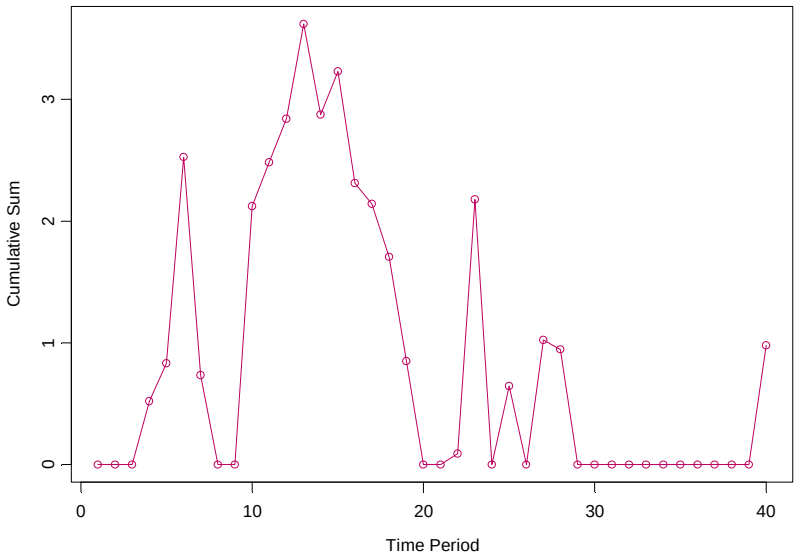
❖ *Region 1: Alarm at 32nd Observation*



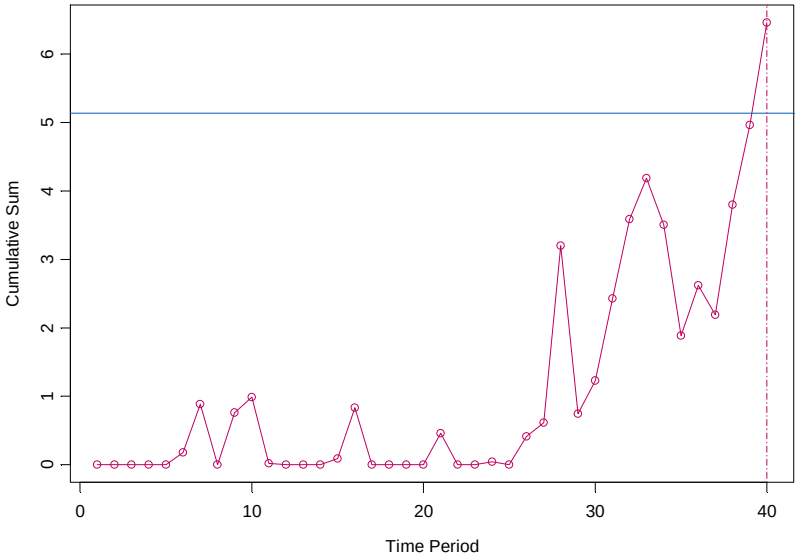
❖ *Region 2: Alarm at 30th Observation*



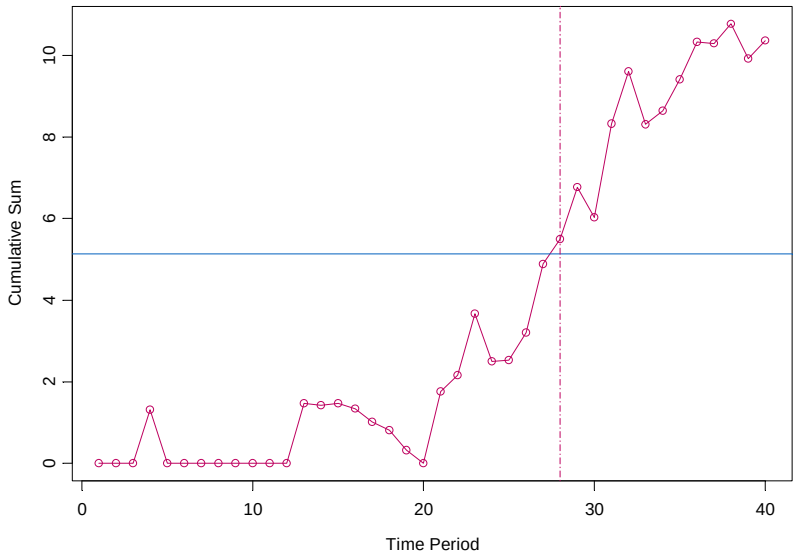
❖ *Region 3: No Alarm*



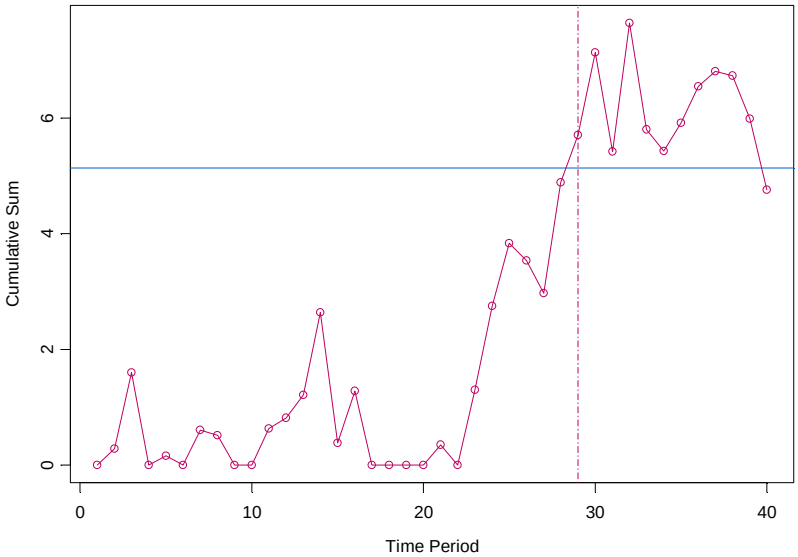
❖ *Region 4: Alarm at 40th Observation*



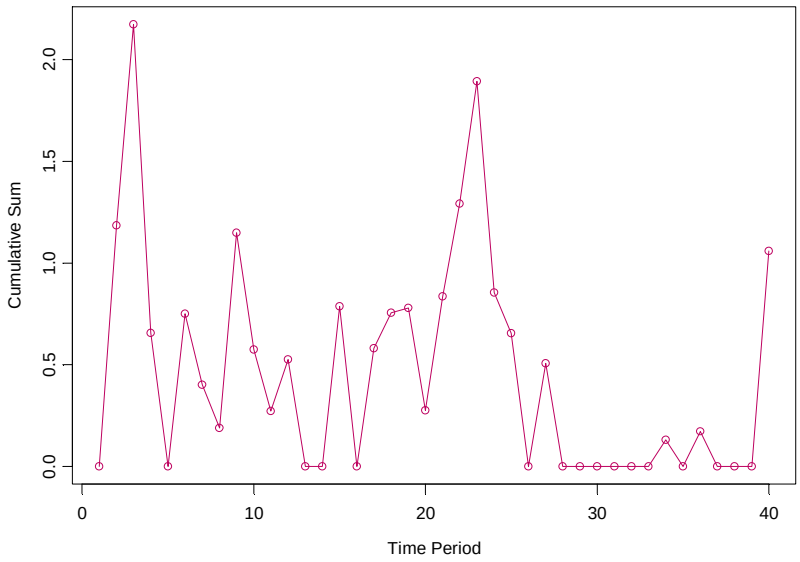
❖ *Region 5: Alarm at 28th Observation*



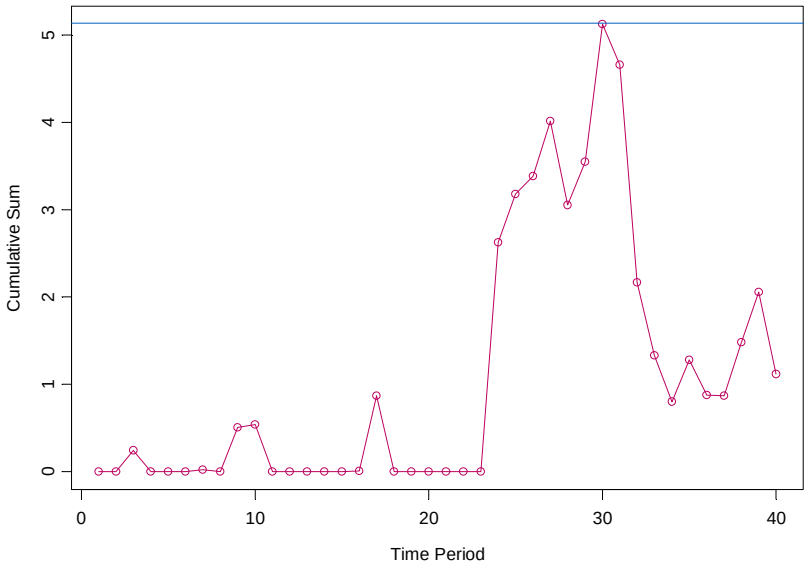
❖ *Region 6: Alarm at 29th Observation*



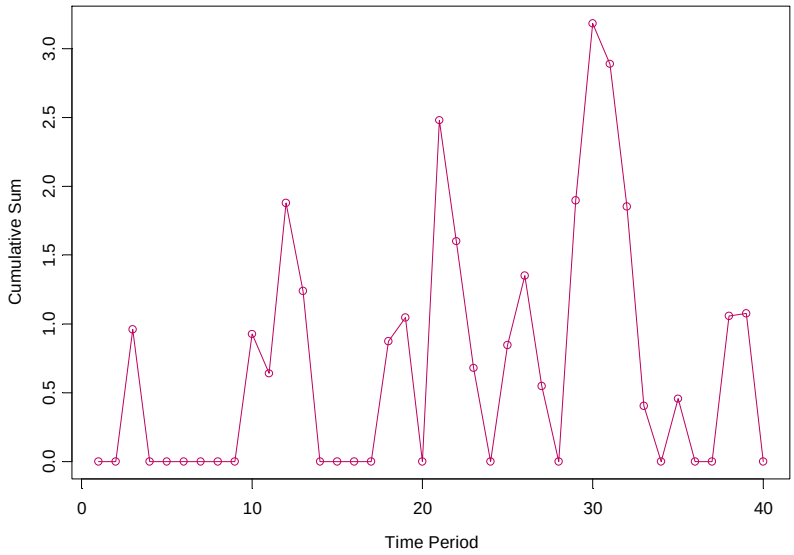
❖ *Region 7: No ALARM*



❖ *Region 8: No Alarm*



❖ *Region 9: No Alarm*



References

1. Åkermo, G. (1994) Constant predictive value of an alarm. *Research Report 1994:6* .Department of Statistics, Göteborg University.

2. Al-Osh, M.A. and Alzaid, A.A. (1987) First-order integer-valued autoregressive (INAR(1)) process. *Journal of Times Series Analysis*, **8**, 261–275.

3. Andersson, E., (2003) A monitoring system for detecting starts and declines of influenza epidemics. *Research Report 2003:12.* Department of Statistics, Göteborg University.

4. Andersson, E., (2004) The impact of intensity in surveillance of cyclical processes. *Communications in Statistics - Simulation and Computation*, **33:4**, 889-913.

5. Arnkelsdottir, H. (1995) Surveillance of rare events: on evaluations of the Sets method. *Research Report 1995:1*. Department of Statistics, Göteborg University.

6. Assunção, R., Correa, T. (2009) Surveillance to detect emerging space-time clusters. *Computational Statistics and Data Analysis*, **53**, 2817-2830.

7. Barbujani,G. (1987) A review of statistical methods for continuous monitoring of malformation frequencies. *Eur.J. Epidem.*, **3**, 67–77.

8. Barbujani, G. and Calzolari, E. (1984) Comparison of two statistical techniques for the surveillance of birth defects through a Monte Carlo simulation. *Statist. Med.*, **3**, 239–247.

9. Bose, A. (1988) Edgeworth Correction by Bootstrap in Autoregressions. *The Annals of Statistics*, **4(Vol. 16)**, 1709-1722.

10. Box, G.E.P. and Jenkins, G.M. (1970). Time Series Analysis: Forecasting and Control. *San Francisco: Holden-Day.*

11. Brockwell, P. J., Davis, R. A. (2002) (Book) Introduction to Time Series and Forecasting(2nd Edition). Springer.

12. Cardinal, M., Roy, R. and Lambert, J. (1999) On the application of integer-valued time series models for the analysis of disease incidence. *Statist. Med.*, **18**, 2025–2039.

13. Chang, T.C. (2008) Cumulative sum schemes for surgical performance monitoring. *J. R. Statist. Soc. A*, **171:2**,407-432.

14. Chen, D., Cunningham, J., Moore, K., Tian,J. (2011) Spatial and temporal aberration detection methods for disease outbreaks in syndromic surveillance systems. *Annals of GIS* , **17:4**, 211-220.

15. Chen, R. (1978) A surveillance system for congenital malformations. *J. Am. Statist. Ass.*, **73**, 323–327.

16. Chen, R. (1983) A monitoring system for chronic diseases-Determining the parameters involved. *Meth. Inform. Med.*, **22**, 149-150.

17. Chen, R. (1986) Revised values for the parameters of the Sets technique for monitoring the incidence rate of a rare disease. *Meth. Inform. Med.*, **25**, 47–49.

18. Chen, R. (1987) The relative efficiency of the sets and the CUSUM techniques in monitoring the occurrence of a rare event. *Statist. Med.*, **6**, 517–525.

19. Chen, R., Froom, P. (2003) The CUSCORE test and the q-interval in cluster analyses of colon cancer and of lymphoma among asbestos workers. *Statist. Med.*, **22**, 3101–3109.

20. Chen, R., Connelly, R. R. and Mantel, N. (1993) Analyzing post-alarm data in a monitoring-system in order to accept or reject the alarm. *Statist. Med.*, **12**, 1807–1812.

21. Chen, R., Iscovich, J. and Goldbourt, U. (1997) Clustering of leukaemia cases in a city in Israel. *Statist. Med.*,**16**, 1873–1887.

22. Chen, R., Mantel, N., Connelly, R. R. and Isacson, P. (1982) A monitoring system for chronic diseases. *Meth.Inform. Med.*, **21**, 86–90.

23. Choi, K. and Thacker, S.B. (1981) An evaluation of influenza mortality surveillance,1962–1979. 1: Time series forecasts of expected pneumonia and influenza deaths. *American Journal of Epidemiology*, **113**, 215–226.

24. Choi ,B.Y, Kim, H., Go, U.Y., Jeong, J.H., Lee,J.W. (2010) Comparison of various statistical methods for detecting disease outbreaks. *Comput. Stat.*, **25**, 603–617.

25. Cooper, B. and Lipsitch, M. (2004) The analysis of hospital infection data using hidden Markov models. *Biostatistics*, **5**, 223–237.

26. Costagliola, D., Flahault, A., Galinec, D., Garnerin, P., Menares, J. and Valleron, A.-J. (1991) A routine tool for detection and assessment of epidemics of influenza-like syndromes in France. *American Journal of Public Health*, **81**, 97–99.

27. Cowling, B.J., Wong, I.O.L, Ho, L.M., Riley,S., Leung, G.M. (2006) Methods for monitoring influenza surveillance data. *International Journal of Epidemiology*, **35**, 1314–1321.

28. Daniel, W.W. (1995) Biostatistics: A Foundation for Analysis in the Health Sciences. *New York: John Wiley & Sons, Inc.*

29. Davison, A.C. and Snell, E.J. (1991) Residuals and diagnostics. In D.V. Hinkley, N. Reid and E.J. Snell (eds), *Statistical Theory and Modelling*. London: Chapman & Hall.

30. De Maré, J. (1980) Optimal prediction of catastrophes with applications to gaussian processes. *The annals of Prob.*,**8**, 841-850.

31. Du, J.G. and Li, Y. (1991) The integer-valued autoregressive (INAR(p)) model. *Journal of Times Series Analysis*, **12**, 129–142.

32. Efron, B. and Tibshirani, R. (1986) Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1(Vol. 1)** ,54-75

33. 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. *J. R. Statist. Soc.* A, **159**, 547–563.

34. Flahault, A., Garnerin, P., Chauvin, P., Farran, N., Saidi, Y., Diaz, C. and Toubiana, L. (1995) Sentinelle traces of an epidemic of acute gastroenteritis in France. *Lancet*, **346**,162–163.

35. Fleming, D.M. and Cohen J.-M. (1996) Experience of European collaboration in influenza surveillance in the winter of 1993-1994. *J. of Public Health Medicine*. **18**,133-142.

36. Frisén, M. (1992) Evaluations of methods for statistical surveillance. *Statist. Med.*, **11**, 1489–1502.

37. Frisén, M. (2003a) Statistical surveillance: optimality and methods. *Int. Stat. Review*, **71:2**, 403-434.

38. Frisén, M. (2003b) Statistical measures for evaluation of methods for syndromic surveillance. *Research Report 2003:11*.Department of Statistics, Göteborg University.

39. Frisén, M. (2007) Properties and Use of the Shewhart Method and Followers. *Sequential Analysis*, **26**, 171-193.

40. Frisén, M. (2009) Optimal Sequential Surveillance for Finance, Public Health, and Other Areas. *Sequential Analysis*, **28:3**, 310-337.

41. Frisén, M. (2010) Discussion on "Quickest Detection Problems: Fifty Years Later" by Albert N. Shiryaev. *Sequential Analysis*, **29**, 394–397.

42. Frisén, M. (2011) Methods and evaluations for surveillance in industry, business, finance and public health. *Qual. Reliab. Eng. Int.*, **27:5**, 611-621.

43. Frisén, M. and Åkermo, G. (1993) Comparison between two methods of surveillance: Exponentially weighted moving average vs CUSUM. *Research Report 1993:1*.Department of Statistics, Göteborg University.

44. Frisén, M. and Cassel, C. (1994) Visual evaluation of statistical surveillance. *Research Report 1994:3*.Department of Statistics, Göteborg University.

45. Frisén, M. and Gottlow, M. (2003) Graphical evaluation of statistical surveillance. *Research Report 2003:10*. Department of Statistics, Göteborg University.

46. Frisén, M. and de Maré, J. (1991) Optimal surveillance. *Biometrika*, **78**, 271–280.

47. Frisén, M. and Sonesson, C. (2002) Optimal surveillance based on exponentially weighted moving averages. *Research Report 2002:1*. Department of Statistics, Göteborg University, Göteborg.

48. Frisén, M. and Wessman, P. (1999) Evaluations of likelihood ratio methods for surveillance: differences and robustness. *Communs Statist. Simuln Computn*, **28**, 597–622.

49. Gallego, B. (2010) Temporal and Spatial Clustering of Bacterial Genotypes. *Infectious Disease Informatics, Springer:New York*, **chapter 18**, 359-371.

50. Gallus, G.,Mandelli, C.,Marchi, M. and Radaelli, G. (1986) On surveillance methods for congenital malformations. *Statist. Med.*, **5**, 565–571.

51. Gallus, G., Radaelli, G. and Marchi, M. (1991) Poisson approximation to a negative binomial process in the surveillance of rare health events. *Meth. Inform. Med.*, **30**, 206–209.

52. German, R. R. (2000) Sensitivity and predictive value positive measurements for public health surveillance systems. *Epidemiology*, **11**, 720–727.

53. Gordon, K. and Smith, A. F. M. (1990) Modeling and monitoring biomedical time series. *J. Am. Statist. Ass.*, **85**,328–337.

54. Han, S.W. (2010) Efficient change detection methods for bio and healthcare surveillance. *PhD Dissertation*, Georgia Institute of Technology.

55. Hawkins, D.M. and Olwell, D.H. (1998) Cumulative Sum Charts and Charting for Quality Improvement. *New York: Springer-Verlag.*

56. Healy, M. J. R. (1983) A simple method for monitoring routine statistics. *Statistician*, **32**, 347–349.

57. Helfenstein, U. (1986) Box–Jenkins modelling of some viral infectious diseases. *Statistics in Medicine*, **5**, 37–47.

58. Hutwagner, L. C., Maloney, E. K., Bean, N. H., Slutsker, L. and Martin, S. M. (1997) Using laboratory-based surveillance data for prevention: an algorithm for detecting Salmonella outbreaks. *Emgng Infect. Dis.*, **3**, 395–400.

59. Järpe, E. (1999) Surveillance of the interaction parameter of the Ising model. *Communs Statist. Theory Meth.*,**28**, 3009–3027.

60. Järpe, E. (2000) Detection of environmental catastrophes. *Research Report 2000:6*. Department of Statistics, Göteborg University.

61. Jiang, W., Han, S.W., Tsui, K.L., Woodall, W.H. (2011) Spatiotemporal surveillance methods in the presence of spatial correlation. *Statist. Med.*, **30**, 569–583.

62. Joner, M.D. Jr., Woodall, W.H., Reynolds, M.R. Jr. (2008) Detecting a rate increase using a Bernoulli scan statistic. *Statist. Med.*, **27**,2555–2575.

63. Johnson, G.D. (2008) Prospective spatial prediction of infectious disease: experience of New York State (USA) with West Nile Virus and proposed directions for improved surveillance. *Environ. Ecol. Stat.*, **15**, 293–311.

64. Jonsson, R. (2010) A CUSUM procedure for detection of outbreaks in Poisson distributed medical health events. *Research Report 2010:4*. Department of Statistics, Göteborg University.

65. Kafadar, K. and Stroup, D. F. (1992) Analysis of aberrations in public health surveillance data: Estimating variances on correlated samples, *Stat. in Medic.*, **12(vol. 11)**, 1551-1568.

66. Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, **82**, 35–45.

67. Kenett, R. and Pollak, M. (1983) On sequential detection of a shift in the probability of a rare event. *J.of the American Stat. Assoc.*, **78**, 389-395.

68. Kjelle, P. E. (1987) Alarm criteria for the fixed gamma radiation monitoring stations. *Report 87-07*. Swedish Radiation Protection Institute, Stockholm.

69. Knoth, S. (2009) Discussion on "Optimal Sequential Surveillance for Finance, Public Health, and Other Areas" by Marianne Frisén. *Sequential Analysis*, **28**, 352–356.

70. Knox, E. G. (1964) The detection of space-time interactions. *Appl. Statist.*, **13**, 25–29.

71. Kulldorff, M. (1997) A spatial scan statistic. *Communs Statist. Theory Meth.*, **26**, 1481–1496.

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

73. Kulldorff, M. (2005) Scan statistics for geographical disease surveillance:an overview, *Lawson A.B. & Kleinman K., Spatial and Syndromic Surveillance for Public Health New York:John Wiley&Sons Ltd.*, **chapter 7,** 115-132.

74. Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., Mostashari, F. (2005) A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS medicine*, **2**, 216-224.

75. Kunsch, H.R. (1989) The Jackknife and the Bootstrap for General Stationary Observations. *The Annals of Statistics*, **3(Vol. 17)**, 1217-1241.

76. Lai, T. L. (1995) Sequential changepoint detection in quality control and dynamical systems (with discussion).*J. R. Statist. Soc.* B, **57**, 613–658.

77. Lai, T. L. (1998) Information bounds and quick detection of parameters in stochastic systems. *IEEE Transactions on Inform. Theory*, **44**, 2917-2929.

78. Lai, T. L. and Shan, Z. (1999) Efficient recursive algorithms for detection of abrupt changes in signals and control systems. *IEEE Transactions on Automatic Control*, **44**, 952-966.

79. Latour, A. (1997) The multivariate GINAR(p) process. *Advances in Applied Probability*, **29**, 228–248.

80. Latour, A. (1998) Existence and stochastic structure of a non-negative integer-valued autoregressive processes. *Journal of Times Series Analysis*, **19**, 439–455.

81. Lawson, A. B. (2001) (Book) Statistical Methods in Spatial Epidemiology. New York: Wiley.

82. Lawson, A. B. and Kleinman, K. (2005) (Book) Spatial and syndromic surveillance in public health, *John Wiley & Sons Ltd.*

83. Le Strat, Y. and Carrat, F. (1999) Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, **18**, 3463–3478.

84. Lie, R. T., Heuch, I. and Irgens, L. M. (1993) A new sequential procedure for surveillance of Down's syndrome. *Statist. Med.*, **12**, 13–25.

85. Limaye , S.S., Mastrangelo, C.M., Zerr, D.M. (2008) A case study in monitoring hospital-associated infections with count control charts. *Quality Engineering*, **20**, 404–413.

86. Liu, R.Y. (1988) Bootstrap Procedures under some Non-I.I.D. Models. *The Annals of Statistics*, **4(Vol. 16)**, 1696-1708.

87. Ljung, G.M. and Box, G.E.P. (1978) On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

88. Lucas, J. M. (1985) Counted data CUSUM's. *Technometrics*, **27**, 129–144.

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

90. Montgomery, D. C., (2009) (Book) Statistical quality control: A modern Introduction (6^{th} Edition). New York: Wiley

91. Moustakides, G. V. (1985) Optimal stopping times for detecting changes in distributions. *Ann. Statist.*, **14**, 1379–1387.

92. Munford, A. G. (1980) A control chart based on cumulative scores. *Appl.Statist.*, **29**,**No 3**, 252-258.

93. Ngo, L., Tager, I. B. and Hadley, D. (1996) Application of exponential smoothing for nosocomial infection surveillance. *Am. J. Epidem.*, **143**, 637–647.

94. Nobre, F. F. and Stroup, D. F. (1994) A monitoring system to detect changes in public health surveillance data. *Int. J. Epidem.*, **23**, 408–418.

95. Nobre, F.F., Monteiro, A.B.S., Telles, P.R. and Williamson, G.D. (2001) Dynamic linear model and SARIMA: a comparison of their forecasting performance in epidemiology. *Statistics in Medicine*, **20**, 3051–3069.

96. Pelecanos, A.M., Ryan, P.A., Gatton, M.L. (2010) Outbreak detection algorithms for seasonal disease data: a case study using ross river virus disease. *BMC Medical Informatics and Decision Making*, **10**.

97. Pollak, M. (1985) Optimal detection of a change in distribution. *Ann. Math. Statist.*, **13**, 206–227.

98. Pollak, M. and Kenett, R. S. (1983) On sequential detection of a shift in the probability of a rare event. *J. Am. Statist. Ass.*, **78**, 389–395.

99. Pollak, M. and Siegmund, D. (1985) A diffusion process and its applications to detecting a change in the drift of Brownian motion. *Biometrica*, **72(2)**, 267-280.

100. Praus, M., Schindel, F., Fescharek, R. and Schwarz, S. (1993) Alert systems for post-marketing surveillance of adverse drug reactions. *Statist. Med.*, **12**, 2383–2393.

101. Quenel, P. and Dab,W. (1998) Influenza A and B epidemic criteria based on time-series analysis of health services surveillance data. *Eur. J. Epidem.*, **14**, 275–285.

102. Radaelli, G. (1992) Using the Cuscore technique in the surveillance of rare health events. *J. Appl. Statist.*, **19**,75–81.

103. Radaelli, R. and Gallus, G. (1989) On detection of a change in the dynamics of rare health events. *Communs Statist. Theory Meth.*, **18**, 579–590.

104. Rath, T.M., Carreras, M. and Sebastiani, P. (2003) Automated detection of influenza epidemics with hidden Markov models., (*M.R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, C. Borgelt and F. Pfenning (eds)), Advances in Intelligent Data Analysis V, 521–531. Berlin: Springer-Verlag.*

105. Raubertas, R. F. (1989) An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statist. Med.*, **8**, 267–271.

106. Reis, B.Y. and Mandl, K.D. (2003) Time series modeling for syndromic surveillance. *BMC Medical Informatics and Decision Making*, **3**, 1–11.

107. Rigau-Perez, J. G., Millard, P. S., Walker, D. R., Deseda, C. C. and Casta-Velez, A. (1999) A deviation bar chart for detecting dengue outbreaks in Puerto Rico. *Am. J. Publ. Hlth*, **89**, 374–378.

108. Ritov, Y. (1990) Decision theoretic optimality of the CUSUM procedure. *Annals of Stat.*, **18**, 1464-1469.

109. Roberts, S.W. (1966) A comparison of some control chart procedures. *Technometrics*, **8**, 411-430.

110. Robertson, C., Nelson, T.A. (2010) Review of software for space-time disease surveillance. *International Journal of Health Geographics*,**9:16**.

111. Rodeiro, C.L.V., Lawson, A.B. (2006) Monitoring changes in spatio-temporal maps of disease. *Biometrical Journal*, **48**, 463–480.

112. Rogerson, P. A. (1997) Surveillance systems for monitoring the development of spatial patterns. *Statist. Med.*, **16**, 2081–2093.

113. Rogerson(2001) Monitoring point patterns for the development of space–time clusters. *J. R. Statist. Soc.* A, **164**, 87–96.

114. Rogerson, P.A. (2005) Spatial surveillance and cumulative sum methods , *Lawson A.B. & Kleinman K., Spatial and Syndromic Surveillance for Public Health New York:John Wiley&Sons Ltd.*, **chapter 6,** 95-114.

115. Rogerson, P. (2006) Formulas for the design of CUSUM quality control charts. Submitted for publication. *Commun. in Stat.-Theory and Methods*, **35(2)**, 373-383.

116. Rogerson, P. and Sun, Y. (2001) Spatial monitoring of geographic patterns: an application to crime analysis. *Computers, Environment, and Urban Systems*, **25**, 539–556.

117. Rogerson, P. and Yamada, I. (2004a) Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, **53**, 79–85.

118. Rogerson, P. and Yamada, I. (2004b) Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches. *Statist. Med.*, **23**, 2195–2214.

119. Rolfhamre, P., Ekdahl, K. (2006) An evaluation and comparison of three commonly used statistical models for automatic detection of outbreaks in

epidemiological data of communicable diseases. *Epidemiol. Infect.*, **134**, 863–871.

120. Rossi, G., Lampugnani, L. and Marchi, M. (1999) An approximate CUSUM procedure for surveillance of health events. *Statist. Med.*, **18**, 2111–2122.

121. Schiöler, L., Frisén, M. (2012) Multivariate outbreak detection. *Journal of Applied Statistics*, **39:2**, 223-242.

122. Schlain, B. R., Lavin, P. T. and Hayden, C. L. (1992) Using an autoregressive model to detect departures from steady states in unequally spaced tumor biomarker data. *Statist. Med.*, **11**, 515–532.

123. Schlain, B. R., Lavin, P. T. and Hayden, C. L. (1993) Using autoregressive and randomwalk models to detect trends and shifts in unequally spaced tumor biomarker data. *Statist. Med.*, **12**, 265–279.

124. Schnell, D., Zaidi, A. and Reynolds, G. (1989) A time series analysis of gonorrhea surveillance data. *Statistics in Medicine*, **8**, 343–352.

125. Sego, L.H. (2006) Applications of Control Charts in Medicine and Epidemiology. *PhD Thesis*, *United States-Virginia, Virginia Polytechnic Institute and State University.*

126. Sego, L.H., Woodall, W.H., Reynolds, M.R. (2008) A comparison of surveillance methods for small incidence rates. *Stat. in Medic.*, **27**, 1225-1247.

127. Serfling, R.E. (1963) Methods for current statistical analysis of excess pneumonia–influenza deaths. *Public Health Reports*, **78**, 494–506.

128. Shiryaev, A. N. (1963) On optimum methods in quickest detection problems. *Theory Probab. Applic.*, **8**, 22–46.

129. Shiryaev (1978) *Optimal Stopping Rules*. New York: Springer.

130. Shore, D. L. and Quade, D. (1989) A surveillance system based on a short memory scheme. *Statist. Med.*, **8**,311–322.

131. Shu, L., Jiang, W., Tsui, K.L. (2010) Weighted CUSUM Procedures for Surveillance of Health Events with Varying Population Sizes. *IEEM* , 467-471.

132. Shu, L., Jiang, W., Tsui, K.L. (2011) A comparison of weighted CUSUM procedures that account for monotone changes in population size. *Statist. Med.*, **30**, 725–741.

133. Siegmund, D. (1985) Sequential Analysis: Tests and Confidence Intervals. *New York:Springer-Verlag*.

134. Siegmund, D. and Venkatraman, E. S. (1995) Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.*, **23**, 255–271.

135. Simonsen, L., Clarke, M. J., Stroup, D. F., Williamson, D., Arden, N. H. and Cox, N. J. (1997) A method for timely assessment of influenza-associated mortality in the United States. *Epidemiology*, **8**, 390–395.

136. Sitter, R. R., Hanrahan, L. P., Demets, D. and Anderson, H. A. (1990) A monitoring system to detect increased rates of cancer incidence. *Am. J. Epidem.*, **132**, 123–130.

137. Smith, A. F. and West, M. (1983) Monitoring renal transplants: an application of the multiprocess Kalman filter. *Biometrics*, **39**, 867–878.

138. Smith, A. F. M., West, M., Gordon, K., Knapp, M. S. and Trimble, I. M.G. (1983) Monitoring kidney transplant patients. *Statistician*, **32**, 46–54.

139. Sonesson, C. (2007) A CUSUM framework for detection of space–time disease clusters using scan statistics. *Statist. Med.*, **26**, 4770–4789.

140. Sonesson, C. and Bock, D. (2003) A review and discussion of prospective statistical surveillance in public health. J. R. Statistic. Soc. A, **166**, Part 1, 5-21.

141. Srivastava, M. S. and Wu, Y. (1993) Comparison of EWMA, CUSUM and Shiryaev-Roberts procedures for detecting a shift in the mean. *Ann. Statist.*, **21**, 645–670.

142. Steiner, S. H., Cook, J. C., Farewell, V. T. and Treasure, T. (2000) Monitoring surgical performance using risk adjusted cumulative sum charts. *Biostatistics*, **1**, 441–452.

143. Steutel, F.W. and van Harn, K. (1979) Discrete analogues of self-decomposability and stability. *Annals of Probability*, **7**, 893–899.

144. Stroup, D. F. and Thacker, S. B. (1993) A bayesian approach to the detection of aberrations in public health surveillance data. *Epidemiology*, **4**, 435–443.

145. Stroup, D. F., Thacker, S. B. and Herndon, J. L. (1988) Application of multiple time-series analysis to the estimation of pneumonia and influenza mortality by age 1962-1983. *Statist. Med.*, **7**, 1045–1059.

146. Stroup,D. F., Wharton, M., Kafadar, K. and Dean, A.G. (1993) Evaluation of a method for detecting aberrations in public health surveillance data. *Am. J. Epidem.*, **137**, 373–380.

147. Stroup,D.F., Williamson,G.D., Herndon, J. L. and Karon, J.M.(1989) Detection of aberrations in the occurrence of notifiable diseases surveillance data. *Statist. Med.*, **8**, 323–329.

148. Stroup, D.F., Williamson, D.F., Dean, A.D., Haddad, S., Basha, M. and Rapose, W. (1994) Statistical Software for Public Health Surveillance. *Atlanta, GA: Centers for Disease Control and Prevention.*

149. Sveréus, A. (1995) Detection of gradual changes: statistical methods in postmarketing surveillance. *Research Report 1995:2.* Department of Statistics, Göteborg University.

150. Takahashi, K. Kulldorff, M., Tango, T., Yih, K. (2008) A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics*, **7:14**.

151. Tango, T. (1995) A class of tests for detecting "general" and "focused" clustering of rare diseases. *Statist. Med.*,**14**, 2323–2334.

152. Tango, T. (2010) (Book) Statistical Methods for Disease Clustering. *Springer: New York.*

153. Testik, M.C. (2007) Conditional and marginal performance of the Poisson CUSUM control chart with parameter estimation. *International Journal of Production Research*, **45:23**, 5621–5638.

154. Tsui, K.L., Han, S.W., Jiang, W., Woodall, W.H. (2011) A Review and Comparison of Likelihood Based Charting Methods., *IIE Transac.*

155. Tsui, K.L., Chiu, W., Gierlich,P., Goldsman, D., Liu, X., Maschek,T. (2008) A Review of Healthcare, Public Health and Syndromic Surveillance. *Quality Engineering*, **20**, 435–450.

156. Unkel, S., Farrington, C.P., Garthwaite, P.H. , Robertson, C., Andrews,N. (2012) Statistical methods for the prospective detection of infectious disease outbreaks: a review. *J. R. Statist. Soc. A*, **175**, 49-82.

157. Vallet, A.C., Lawson, A.B. (2011) Conditional predictive inference for online surveillance of spatial disease incidence. *Statist. Med.*, **30**,3095–3116.

158. Vanbrackle, L. and Williamson, G. D. (1999) A study of the average run length characteristics of the National Notifiable Diseases Surveillance System. *Statist. Med.*, **18**, 3309–3319.

159. Waller, L.A. and Gotway, C.A. (2004) (Book) Applied spatial statistics for public health data. *John Wiley*.

160. Watier, L., Richardson, S. and Hubert, B. (1991) A time-series construction of an alert threshold with application to S-Bovismorbificans in France. *Statist. Med.*, **10**, 1493–1509.

161. West, M. and Harrison, J. (1989) Bayesian Forecasting and Dynamic Models. *New York: Springer-Verlag.*

162. Wharton, M., Price, W., Hoesly, F., Woolard, D., White, K., Greene, C. and McNabb, S. (1993) Evaluation of a method for detecting outbreaks of diseases in 6 states. *Am. J. Prev. Med.*, **9**, 45–49.

163. Whittaker, J. and Frühwirth-Schnatter, S. (1994) A dynamic changepoint model for detecting the onset of growth in bacteriological infections. *Appl. Statist.*, **43**, 625–640.

164. Williamson, G. D. and Hudson, G. W. (1999) A monitoring system for detecting aberrations in public health surveillance reports. *Statist. Med.*, **18**, 3283–3298.

165. Wolter, C. (1987) Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's Sets technique. *Meth. Inform. Med.*, **26**, 215–219.

166. Woodall, W. H. (1997) Control charts based on attribute data: bibliography and review. *J. Qual. Technol.*, **29**,172–183.

167. Woodall, W.H. (2006) The use of control charts in health-care and public-health surveillance. *Journal of Qual. And Techn.*, **38:2**, 89-104.

168. Woodall, W.H., Marshall, J.B., Joner, M.D., Fraker, S.E., Abdel-Salam, G. (2008) On the use and evaluation of prospective scan methods for health-related surveillance. *J. R. Statist. Soc. A*, **171:1**, 223-237.

169. Wu, Z., Jiao,J., He,Z. (2009) A single control chart for monitoring the frequency and magnitude of an event. *Int. J. Production Economics*, **119**,24–33.

170. Yann L.S. (2005) Overview of temporal surveillance, *Lawson A.B. & Kleinman K., Spatial and Syndromic Surveillance for Public Health New York:John Wiley&Sons Ltd.,* **chapter 2,** 13-29.

171. Yashchin, E. (1993) Statistical control schemes-Methods, applications and generalizations. *Intern. Stat. Review*, **61**, 41-66.

172. Zaidi, A., Schnell, D. and Reynolds, G. (1989) Time series analysis of syphilis surveillance data. *Statistics in Medicine*, **8**, 353–362.

173. Zeger, S.L. (1988) A regression model for time series of counts. *Biometrika*, **75**, 621–629.