Space-time surveillance for the detection of emerging clusters

Renato Assunção[1], **Thaís Correa**[1]

[1] Laboratório de Estatística Espacial (LESTE)
Depatamento de Estatística – Universidade Federal de Minas Gerais (UFMG)
Av. Antônio Carlos 6627 – 31270-901 – Belo Horizonte – MG – Brazil

assuncao@est.ufmg.br, tataest@yahoo.com.br

Abstract. *Space-time events with coordinates (x_i, y_i, t_i) are monitored continually. The events' density varies largely in both, space and time. At a certain unknown instant τ, a relatively small cluster of increased intensity starts to emerge. Its location is also unknown. The aim is to let make an alarm go off as soon as possible after τ but avoiding it to go off unnecessarily. In this paper we propose an alarm system that does not require the specification of the spatial pattern or the temporal pattern. It is based on a martingale approach. We detail its theoretical foundation and the corresponding algorithm. We provide an illustration of its use in practice.*

Resumo. *Eventos espaço-temporais com coordenadas (x_i, y_i, t_i) são monitorados continuamente. A densidade dos eventos é bastante variável, tanto no espaço quanto no tempo. Em certo momento desconhecido τ, um cluster relativamente pequeno começa a emergir. Sua localização é desconhecida. O objetivo é fazer um alarme soar logo após τ mas evitando que ele soe desnecessariamente. Neste artigo, nós propomos um sistema alarme que não requer a especificação do padrão espacial ou temporal. Ele é baseado num método de martingalas. Nós damos os detalhes teóricos e o algoritmo correspondente. Nós também fornecemos uma ilustraçã o uso prático do sistema.*

1. Introduction

We are interested in monitoring incoming space-time events to detect, as early as possible, an emergent space-time cluster. Assume that point process events (x_i, y_i, t_i) are continuously recorded where (x_i, y_i) are the spatial coordinates and t_i is occurrence time of the i-th event. At a certain unknown instant τ, a relatively small cluster of increased intensity starts to emerge. Its location is also unknown. The aim is to let make an alarm go off as soon as possible after τ. The alarm system should also provide an estimate of the cluster location. The alarm system should take into account purely spatial and purely temporal heterogeneity.

In this work we propose a space-time surveillance system with these specifications. It does not require the specification of the spatial pattern or the temporal pattern. It is based on a martingale approach. We detail its theoretical foundation and the corresponding algorithm. Due to lack of space, we study its efficiency in another paper.

Epidemiological surveillance systems include early statistical warning methods that aim to provide information which can be acted upon to help in the prevention and control of diseases. There is a renewed interest on the development of statistical systems that include spatially referenced information sources due, among other reasons, to heightened concerns about bioterrorism.

The requirements of a surveillance system accounting for spatial structure are generally structured around a basic trade-off: the need for quickly detecting possible outbreaks and epidemics must be balanced against the need for not triggering alarm signals too often unnecessarily.

In this paper, we describe a method to analyze space and time surveillance data in the form of point processes. We propose a probability model to describe eventually emerging spatial clusters with a minimum requirement of user-defined parameters. Based on this model for the emerging spatial clusters, we use the Shiryaev-Roberts statistic and adopt a martingale approach to derive the test properties. Hence, we are able to control the average length run of our surveillance method under the absence of emerging spatial clusters. We define appropriately the average run length for the situation when there are clusters present in the data and illustrate the method in practice. The algorithm is implemented in a freely available stand-alone software and it is expected soon to be in TERRAVIEW.

2. Literature Review

The traditional methods for space-time cluster detection are retrospective in nature. That is, they search in a a database of past events for evidence of clusters' presence. In contrast, our interest is on prospective methods: an events' database is updated regularly and then an algorithm should run to help deciding on the emergence of localized space-time clusters. Hence, the clusters must be alive in the sense that at least some of the most recent events belong to the eventually detected clusters. This brings several difficult problems well known in the artificial intelligence literature: repeated significance tests (at least one every time the database is updated); trade-off between setting up the system to go off as soon as possible after a localized space-time cluster starts to emerge and, at the same time, requiring that the false alarms frequency be kept at a minimum.

A thorough literature review can be found in the book edited by [A B Lawson 2005] or in [Sonesson et al. 2003]. We give here a very brief overview of the main proposals. There are non-spatial methods derived from quality control ideas concerned with monitoring a stochastic process on time. The Shewart Chart Control is a very simple and popular method but it is not sensitive to small changes in the process. The Cummulative Sum (CUSUM) method accumulates the recent evidence to the previous data to trigger a threshold limit. It has been shown that it has optimal properties in very simple scenarios. Exponentially weighted moving average also accumulates evidence, as the CUSUM method, but it discounts observations as they get old. All these methods assume data are independent in time, not a realistic assumption. [Kennett and Pollak 1996] uses a Shiryaev-Roberts statistics to allow for dependent data.

There are few space-time oriented proposals. Two recent and promising ones are [Kulldorff 2001] who proposed a space-time scan statistic for areal data. [Rogerson 2001] suggested a statistic based on local Knox statistic.

We introduced a new method focusing on point process data. That is, there is no risk population info. The null hypothesis of interest is that we have a separable events density with unspecified and arbitrary spatial and temporal heterogeneity. As alternative, we assume that somewhere, at some moment, few localized space-time high intensity clusters start to emerge. We develop a likelihood model for this pair of hypotheses and monitor the incoming data with a spatial version of the Shyriaev-Roberts statistic.

3. Basic Concepts and notation

The Shiryayev-Roberts Method was developed for temporal processes only. Suppose that a sequence of possibly dependent random variables $X_1, X_2, \dots$ is observed. Let $f_{(k)}(x_1, x_2, \dots x_n)$ be the joint density distribution of the first n random variables when a cluster starts to emerge at moment $\tau = k$. When no cluster ever emerges, we write $f_\infty(x_1, x_2, \dots x_n)$. Any surveillance method implies a stopping time N, the first moment when the alarm goes off. $E_k(\cdot)$ is the expectation with respect to $f_{(k)}$ and $E_\infty(N)$ is called the Average Run Length and it is denoted by ARL^0. Clearly, it is desirable to keep ARL^0 small and, for that, the user establishes an acceptable minimum threshold B for this parameter. That is, we want $ARL^0 = E_\infty(N) > B$. The Shiryayev-Roberts test statistic is given by

$$R_n = \sum_{k=1}^{n} \frac{f_{(k)}(X_1, X_2, \dots, X_n)}{f_{(\infty)}(X_1, X_2, \dots, X_n)}$$

The alarm goes off if R_n is too large, that is, if $R_n \geq A$. The stopping time is N_A: the alarm goes off by the first time at N_A where

$$N_A = \min\left[n \,| R_n \geq A\right]$$

It remains to find A such that $ARL^0 = E_\infty(N) > B$.

Under P_∞, the sequence

$$\Lambda_{k,n} = \frac{f_{(k)}(X_1, X_2, \dots, X_n)}{f_{(\infty)}(X_1, X_2, \dots, X_n)}$$

is a martingale with expected value equal to 1 (even with dependent observations). Therefore, $R_n - n = \sum_{k=1}^{n}(\Lambda_{k,n} - 1)$ is a zero mean martingale. By the Optional Sampling Theorem, we have

$$E_\infty(R_{N_A} - N_A) = 0 \Rightarrow E_\infty(N_A) = E_\infty(R_{N_A}).$$

By definition, $R_{N_A} \geq A$ and hence $E_\infty(N_A) \geq A$. Therefore, taking $A = B$ satisfies the condition $E_\infty(N_B) \geq B$.

There are several advantages associated with the Shiryayev-Roberts (SR) method. First, it can be shown that it exhibits some optimal properties in some simple scenarios. Furthermore, in terms of the delay time for the alarm going off after the purely temporal clusters strats to emerge, the SR and CUSUM are similar. The SR method does not require independence between observations. And it can also be shown that SR is at least as efficient as some optimal classical procedures.

The major disadvantage of the SR method is that it depends on the complete specification of the joint distribution of $X_1, \dots, X_n$ after a change occurs at $\tau = k$. If this is difficult to be done in the purely temporal context, in the space-time situation it seems hopeless. However, we found a way out as we explain next.

4. Our proposal for space-time clusters

Let N be a Poisson process in $\mathbb{R}^3$ partially observed in the three-dimensional region $\mathcal{A} \times [0, T]$. Let $N(C_i)$ be the number of events in the cylinder C_i. $N(C_i) \sim$ Poisson $(\mu(C_i))$ and $\mu(C_i)$ is unknown. Let $\lambda(x, y, t)$ be the intensity function of the events in $\mathcal{A} \times [0, T]$. Consider a cylinder C_i in $\mathbb{R}^3$ (see Figure 1) and let $\mu(C_i)$ be the integral over C_i of $\lambda(x, y, t)$, while μ is the expected number of events in all the region $\mathcal{A} \times [0, T]$. Define the marginal spatial an d temporal densities by $\lambda_S(x, y) = \mu^{-1} \int_{[0,T]} \lambda(x, y, t)\, dt$ and $\lambda_T(t) = \mu^{-1} \int_{\mathcal{A}} \lambda(x, y, t)\, dx\, dy$, respectively.

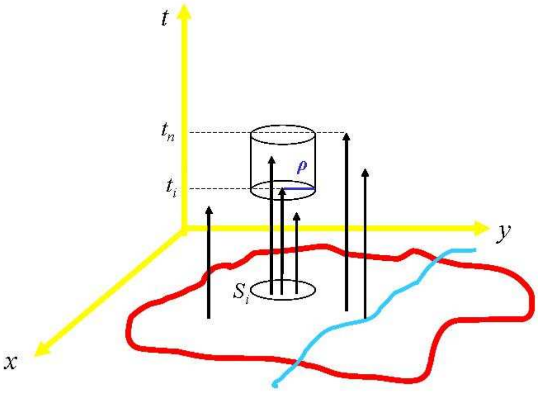

Figure 1. A typical figure

We define now the pair of hypotheses. The null hypothesis (no cluster scenario) is established as a separable intensity $\lambda(x, y, t) = \mu\, \lambda_S(x, y)\, \lambda_T(t)$ where $\lambda_S(x, y)$ and $\lambda_T(t)$ are arbitrary and unspecified. That is, they are nuisance parameters. The alternative hypothesis assumes that there exists a time τ, a constant $\varepsilon > 0$, and a cylinder C_τ (yet to be defined) such that

$$\lambda(x, y, t) = \mu\, \lambda_S(x, y)\, \lambda_T(t)\, (1 + \varepsilon\, I_C(x, y, t))$$

. The parameter ε is the *relative* change on the events intensity within the cluster and it *must* be specified by the user.

TO define a useful class of cylinders C_τ, we start considering that,if a higher incidence cluster emerges, we must be able to detect it through the observed events. That is, *non-events* (or void spaces) do not bring information about an emerging cluster. Hence, we decided to constrain τ to be equal to one of the observed t_i's; the cylinders should be in the form of a circle S times a temporal interval. The time interval is $[t_i, t_n]$ where t_n is the last event, since interest is only in *alive clusters*. The cylinder S has a radius ρ *specified by the user.*

We can now proceed to determine the mean $\mu(C_i)$. From the non-homogeneous Poisson process properties, under the null hypothesis, we have:

$$\mu(C_i) = \int_{C_i} \lambda(x, y, t)\, dx\, dy\, dt = \mu \int_{S_i} \lambda_S(x, y)\, dx\, dy \int_{[t_i, t_n]} \lambda_T(t)\, dt$$

An estimate of $\mu(C_i)$ under H_0 is given by

$$\hat{\mu}(C_i) = \frac{N(S_i \times [0, T])\, N(\mathcal{A} \times [t_i, t_n])}{n}$$

where $N(S_i \times [0,T])$ is the number of events within circle S_i irrespective of time; $N(\mathcal{A} \times [t_i, t_n])$ is the number of events with times between t_i and t_n, irrespective of spatial location; and n is the total number of events (see Figure 2).

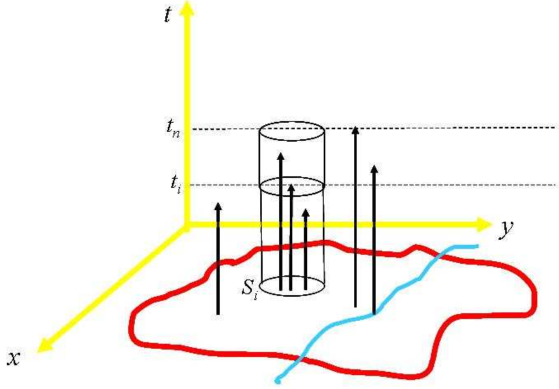

Figure 2. The estimate $\hat{\mu}(C_i)$

To define the test statistic, we consider the likelihood of space-time Poisson processes. Under H_0, we have

$$L_\infty = \left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i)\right) \exp\left(-\int_{R^3} \lambda(x,y,t)\, dx\, dy\, dt\right)$$

Under the alternative, we have

$$\begin{aligned} L_\tau &= \left(\prod_{i=1}^{n} \lambda(x_i, y_i, t_i)\,(1 + \varepsilon\, I_{C_\tau}(x_i, y_i, t_i)\right) \\ &\quad \exp\left(-\int_{R^3} \lambda(x,y,t)\, dx\, dy\, dt\right) \exp\left(-\varepsilon \int_{C_\tau} \lambda(x,y,t)\, dx\, dy\, dt\right) \end{aligned}$$

where $\lambda(x,y,t) = \mu\, \lambda_S(x,y)\, \lambda_T(t)$ and C_τ is the putative cluster cylinder.

Therefore, a space-time version of the SR test statistic R_n becomes

$$\begin{aligned} R_n &= \sum_{\tau=1}^{n} \frac{L_\tau}{L_\infty} \\ &= \sum_{\tau=1}^{n} \left\{ \left[\prod_{i=1}^{n} (1 + \varepsilon\, I_{C_\tau}(x_i, y_i, t_i)\right] \exp\left(-\varepsilon \int_{C_\tau} \lambda(x,y,t)\, dx\, dy\, dt\right) \right\} \\ &= \sum_{\tau=1}^{n} (1+\varepsilon)^{N(C_\tau)} \exp(-\varepsilon\, \mu(C_\tau)) \end{aligned}$$

with $\mu(C_\tau)$ estimated as explained before.

The parameter $\varepsilon > 0$ is known (user-specified) and measures the anticipated relative change in the events' density. Our surveillance method calculates R_{n+1} as the $(n+1)$-th event arrives with with $\hat{\mu}(C_\tau)$ rather than $\mu(C_\tau)$. The alarm goes off when $R_n \geq A$ for the first time. In summary, the algorithm associated with our proposal needs as input: n cases events given by the coordinates x, y and time t; the value of three tuning parameters: ε, the anticipated relative change in density within the cluster; the anticipated radius ρ for the cluster; the threshold A, which should be approximately equal to the desired ARL^0. Iteratively in n, calculate R_n. The output is a sequence of values R_n where n is the number of events. If $R_n > A$ for any n, the alarm goes off.

5. Illustration

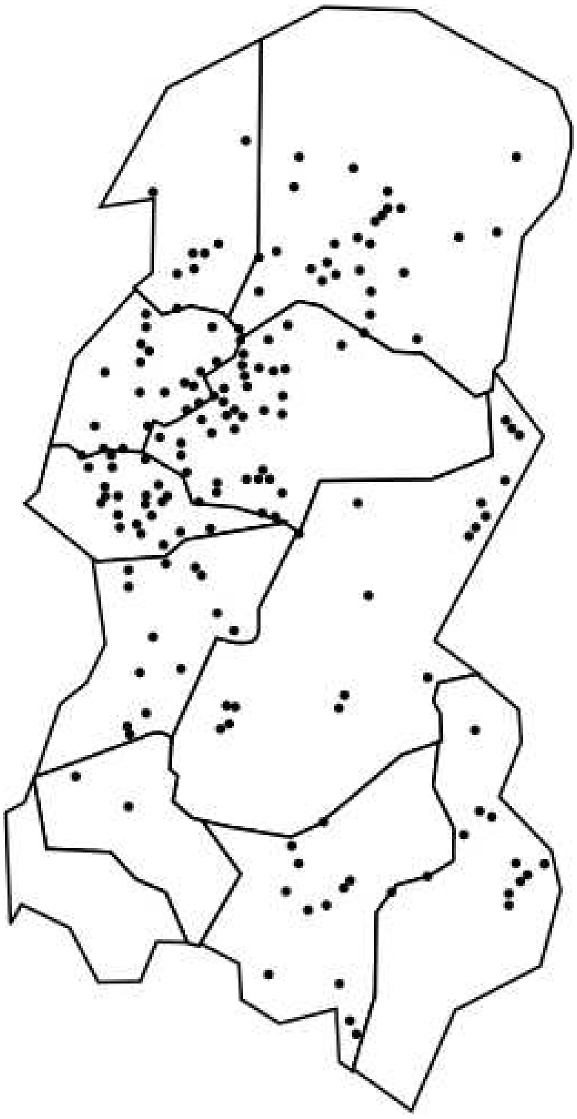

Figure 3. Burkitt lymphoma cases in Uganda

As an illustration, we used a classical example of retrospective detection of space-time clustering: the data with place of residence and onset time for all 188 cases of Burkitt lymphoma between 1961 and 1975 in the West Nile district in Uganda (see Figure 3). [Rogerson 2001] found evidence of space-time clusters using local Knox tests and adopting a probability of false alarm of 0.1. However, we have not been able to reproduce his results using his methods. Apparently, his formulas or graphs with his results are not correct.

The tuning parameters in our surveillance method were:

- We fix $\varepsilon = 0.5$, a large anticipated change.
- $\rho = 210$ km (weighted average of values used by Rogerson (2001)).
- Limit A of the alarm = 161. In average, we expect 161 events before the alarm goes off without need.

Figure 4 shows the R_n versus n. We can observe that the alarm goes off at event number 148 (February, 1973). Typically, there was little variation of the detected space-time cluster over many different tuning parameter choices. One pattern we found is that, for $\rho = 2.5, 5, 10, 20$ km, the smaller ε, the longer it takes for the alarm to go off.

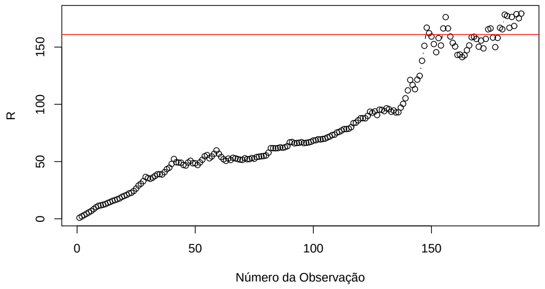

Figure 4. R_n **for** $\varepsilon = 0.5$ **e** $\rho = 20$ **km**

6. Conclusions

Our method has many desirable features. First, it does not require data on the population at risk data, only cases are necessary. Second, it adjusts for purely spatial and purely temporal clustering, and it provides statistical inference for the emerging cluster detected. Third, it does not require many input parameters. We think it will be of great use in many practical applications.

References

A B Lawson, K. K. (2005). *Spatial and Syndromic Surveillance for Public Health*. John Wiley & Sons, New York.

Kennett, R. and Pollak, M. (1996). Data-analytic aspects of the shiryayev-roberts control chart: surveillance of a non-homogeneos poisson process. *Journal of Applied Statistics*, 23:125–137.

Kulldorff, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistc. *Journal Royal Statistical Society*, 164, Part 1:61–72.

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

Sonesson, C., , and Bock, D. (2003). A review and discussion of prospective statistical surveillance in public health. *Journal Royal Statistical Society*, 166, Part 1:5–21.