BIOMETRICS 57, 577–583
June 2001

Multiple Temporal Cluster Detection

Nicolas Molinari,* Chistophe Bonaldi, and Jean-Pierre Daurès

Laboratoire de Biostatistique, Institut Universitaire de Recherche Clinique,
641 Avenue Gaston Giraud, 34093 Montpellier, France
**email:* molinari@helios.ensam.inra.fr

SUMMARY. This article proposes a simple method to determine single or multiple temporal clustering on a variable size population. By a transformation of the data set, the method based on a regression model allows consideration of a variable population size during the time of study. A model selection procedure and a resampling method are used to select the number of clusters. The results have applications in epidemiological studies of rare diseases.

KEY WORDS: Model selection; Nearest neighbor; Regression method; Resampling method.

1. Introduction

In epidemiological studies, when the etiology of a disease has not yet been well established, it is sometimes required to examine data for obtaining evidence of temporal clustering or of cyclical clustering, as in seasonal variations. Let $X_1, \ldots, X_N$ be independent and identically distributed (i.i.d.) random variables that denote the times of occurrence of N events in an interval $(0, T)$. We wish to test the null hypothesis that the events are uniformly distributed against the alternative that they cluster within some subintervals of $(0, T)$.

Ederer, Myers, and Mantel (1964) developed a test for temporal clustering using a cell-occupancy approach. They divided the time period into disjoint subintervals. This test statistic is simply the number of cases occurring in a subinterval. Under the no-clustering hypothesis, the N cases are randomly distributed among the subintervals. However, the resulting chi-square test (used to test the multinomial distribution) does not yield an efficient method.

Consider the following hypothetical example. Suppose we observe a rare disease during 1 year in a little town, say Clusterville. The number of known events is $N = 42$ for the whole year. The study starts the first day and the first event occurs on the 11th day. Every 10 days, we observe another event except from day 181 to day 241, when one event occurs every 5 days. The study stops at day 365. It is clear that $[181, 241]$ is a time window with clustered events. Tango (1984) proposes a test of temporal clustering based on the distribution of counts in disjoint equal time intervals. Whittemore and Keller (1986) showed that the distribution of Tango's index is asymptotically normal. Applying this procedure to Clusterville, assuming six time intervals does not allow rejection of the null hypothesis of uniformity ($p = 0.2$). Note that the cluster interval $[181, 241]$ does not match the interval using Tango's index, which does not contain the event occurring at day 181.

Naus (1965) introduced a test known as the scan test. The test statistic, the maximum number of cases observed in an interval of length t, is found by scaling all intervals of length t in the time period. Statistical significance of the scan test is assessed by using tables of p-values computed by Naus (1966) and Wallenstein (1980) for selected interval lengths, time lengths, and sample sizes. Weinstock (1981) proposed a generalization of the scan test that is adjusted for changes among the population at risk. Unfortunately, with the simulated example, the scan test does not provide a significant statistic to reject the uniformity hypothesis (with six subintervals, $p = 0.38$).

An efficient method for detecting temporal clustering is proposed by Kulldorff and Nagarwalla (1995). With the scan statistics with variable window, the cluster time window size does not need to be chosen *a priori*. This test is the generalized likelihood ratio test for a uniform null distribution against an alternative of nonrandom clustering. Bootstrapped simulations are performed to carry out the significance test. For the example, we obtain $p = 0.09$ with 1000 simulations, and we fixed the minimal number of points at five. The test only considers clusters that contain five or more points (Nagarwalla, 1996). An extension of this method is presented by Kulldorff (1997). The scan statistic with a variable window is used for detecting disease clusters in heterogeneous populations. He introduced a spatial scan statistic for the detection of clusters not explained by the baseline process in heterogeneous populations.

Larsen, Holmes, and Heath (1973) developed a rank-order procedure. The time period is divided into disjoint subintervals that are numbered sequentially. The test statistic is the sum of absolute differences between the rank of the subinterval in which a case occurred and the median subinterval rank. This test is sensitive only to unimodal clustering and cannot distinguish between multiple clustering and randomness.

Huntington and Naus (1975), Cressie (1977), and Hwang (1977) derived and then Naus (1982) accurately approximated the probability of at least one cluster. Barton and David (1956) found the distribution of the number of clusters of size two. McClure (1976) obtained asymptotic results for the distribution of the number of clusters of a given size. Glaz and Naus (1983) established the expectation, variance, and approximate distribution of the number of clusters of a given size. So, with rare diseases, a long time of study is necessary to examine data for evidence of temporal clustering. The problem is that, in this case, the population at risk evolves during time. Due to a natural increasing or to a seasonal evolution, the population at risk is not constant during the time of study.

In the next section, we present a new method for determining data clustering. Based on a simple transformation of the data, our method determines a time window with excess events and, for any position of the window, it scans continuously across the period of observation. Moreover, the method is effective with changes in the population at risk. Existence of one or more clusters is determined by using bootstrapped simulations and a classical model selection procedure. The regression method is explored using simulations that allow for an examination of its properties and also on the classical Knox data set. Another data set consists of 62 spontaneous hemoptysis admissions (pulmonary disease) at Nice hospital from January 1 to December 31, 1995. Detecting periods of significant cluster occurrences brings precious information on the disease. The purpose of this investigation is to adapt conditions of admission or treatment of predisposed patients during a favorable period. Another objective is to point out potential climatic factors, like temperature or hydrometry, that influence the disease occurrence. Nevertheless, since Nice is situated in the south of France, each summer, a lot of tourists increase the population at risk. An estimation of this population is used in our model for detecting clusters.

2. Method Presentation

The approach is first based on a transformation of the data set in order to produce values corresponding to the time (the distance) between two successive events. Under the no-clustering hypothesis, these values can be estimated by a constant, i.e., the mean distance. On the contrary, a piecewise constant model improves the fitting. A classical criterion for selecting models allows determination of the presence of clusters. Statistical tests for cluster detection must have a correct nominal α level. Since the proposed method is not a conventional statistical test, we propose using bootstrapped samples to obtain a p-value and to compare its performance with those of existing statistical tests. At the end of this section, we propose a simple transformation of the data set that considers changes in the population at risk.

2.1 *Data Transformation*

Let $X_1, \ldots, X_N$ be defined as in the Introduction. Without loss of generality, set $T = 1$ throughout this section. Suppose that $x_1, \ldots, x_N$ are dropped at random in the unit interval $(0, 1)$. As indicated in Figure 1, denote the ordered distances of these points from the origin by $x_{(i)}$ $(i = 1, \ldots, N)$ and set $y_i = x_{(i)} - x_{(i-1)}$ $(x_{(0)} = 0)$. Assuming that the X_i's are i.i.d. uniform U(0, 1), the random variables $X_{(1)}, \ldots, X_{(N)}$ are then distributed as N-order statistics from a uniform U(0, 1)

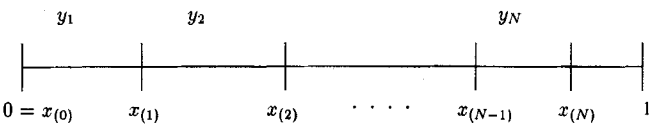

Figure 1. Random division of an interval.

parent, i.e., $X_{(i)}$ follows a beta distribution $\beta(i, N - i + 1)$ and $Y_i = X_{(i)} - X_{(i-1)}$ has a beta distribution $\beta(1, N)$ (see David, 1980). A slightly efficient method for detecting nonrandom clusters of points on a line is, e.g., by verifying, using a Kolmogorov–Smirnov test, that the Y_i's have a beta distribution $\beta(1, N)$. This method is equivalent to testing the assumption of uniformity of the X_i's. In the case where a cluster is present, the test does not provide the time window with excess events. In the next section, we present our method for detecting, according to the Y_i values, the cluster's presence and also for determining its ranges.

2.2 *Data Fitting*

Let $(x_1, \ldots, x_N)$ be a sample of X and $(y_1, \ldots, y_N)$ be the corresponding sample of $Y = (Y_1, \ldots, Y_N)$ defined as in the previous section. Consider the data set $(i, y_i)_{i=1,\ldots,N}$. Under the no-clusters hypothesis, an appropriate regression on this data set is the constant function

$$f(i) = \bar{y} = \frac{1}{N} \sum_{j=1}^{N} y_j. \tag{1}$$

Figure 2 presents the regression function and the data points corresponding to the Clusterville example. Assume that events $x_{(k)}, \ldots, x_{(k+l)}$ are clustered, i.e.,

$$\bar{y}_c = \frac{1}{l} \sum_{i=k+1}^{k+l} y_i < \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{y}.$$

In this case, a better regression model than (1) is the stepwise regression function

$$f(i) = \bar{y}_{\bar{c}} \times \mathrm{I}_{[1;k] \cup [k+l+1;N]}(i) + \bar{y}_c \times \mathrm{I}_{[k+1;k+l]}(i), \tag{2}$$

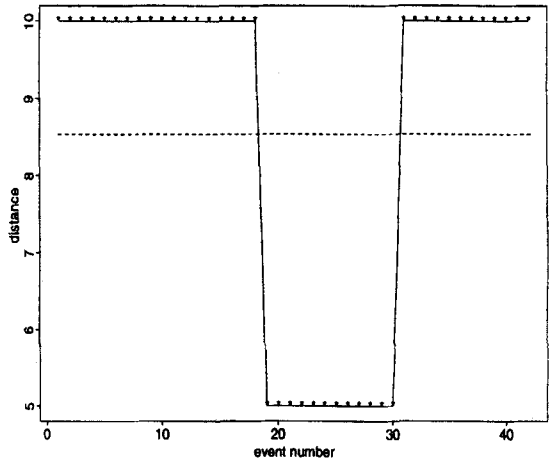

Figure 2. Clusterville data transformation, constant regression (dashed line) and one-cluster regression function (solid line).

where $\bar{y}_{\bar{c}}$ is the mean of the y_i's without $\{y_{k+1}, \ldots, y_{k+l}\}$, the mean of the distance between the nonclustering events, and $I_A(x) = 1$ if $x \in A$ and equals zero if not.

Moreover, with the same approach, the regression model to determine n clusters is

$$f(i) = \bar{y}_{\overline{c_1 \cup \cdots \cup c_n}} \times I_U(i) + \bar{y}_{c_1} \times I_{\mathcal{C}_1}(i) + \cdots + \bar{y}_{c_n} \times I_{\mathcal{C}_n}(i), \quad (3)$$

where $U = [1; k_1] \cup [k_1 + l_1 + 1; k_2] \cup \cdots \cup [k_n + l_n + 1; N] = \overline{\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_n}$ is the uniform repartition range and $\mathcal{C}_p = [k_p + 1; k_p + l_p]$ for $p \in \{1, \ldots, n\}$ are the clustering ranges. In order to compute the bounds $k_1, l_1, \ldots, k_n, l_n$ of each cluster, one needs to solve the classical least squares problem,

$$\min_{k_1, l_1, \ldots, k_n, l_n} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(i))^2. \quad (4)$$

Figure 2 shows the regression function (2) on the Clusterville data set and clearly indicates the presence of a cluster between the 19th and 30th events.

To determine the presence of clusters, to accept the uniformity hypothesis, and also to select the number n of clusters, we use classical criteria, i.e., the Akaike information criteria (AIC; Akaike, 1974) or the Bayesian information criteria (BIC; Schwartz, 1978). Each different number of supposed clusters corresponds to a different regression model with a different criterion value. For example, the BIC corresponding to a constant (no cluster) regression function on the Clusterville data set is 72.2, against $-\infty$ for the function corresponding to the single-cluster model (2).

We apply the method to obtain several models with different numbers of clusters. Once k and l, the cluster bounds, are computed for each model, a simple approach to determining the number of clusters is to select the model with the smallest criterion value. However, to avoid sample effects and to compare the method with other statistical tests, we compute again the criterion on a 1000 bootstrapped samples of $(i, y_i)_{i=1,\ldots,n}$. Let $CRIT_j^i$ and $CRIT_j^{i'}$ be the respective values of the criterion for the i cluster model and the i' $(i < i')$ cluster model for sample j. The value

$$\alpha_{i,i'}^{CRIT} = \frac{1}{1000} \sum_{j=1}^{1000} I\left(CRIT_j^i < CRIT_j^{i'}\right)$$

gives an idea about the α level (nominal level of a conventional statistical test) of the proposed method. For example, $\alpha_{0,1}^{CRIT}$ is the percentage of bootstrapped samples for which the one-cluster model is selected with the criterion $CRIT$ against the no-clustering model. Classically, $\alpha_{i,i'}^{CRIT} < 0.05$ is considered as significant and the i cluster model is chosen against the i' cluster one. The criterion choice (AIC, BIC, or other penalization) is debated in Section 3.1.1.

In the next section, we describe the difficulties of testing nonrandom clustering when the population at risk varies during the time period. A simple modification of the data transformation allows solving this problem, and the method presented above provides an efficient solution.

2.3 *Variable Population Size Effects*

Suppose now that Clusterville is situated near a beautiful beach. Each summer (day 182 to day 243 for July 1 and August 31), the local population accommodates a lot of tourists and, as a consequence, the number of inhabitant doubles in this period. The number of the rare disease events follows the same rule and the data set presented in the Introduction is uniform.

During the studied period, the number of people who should be affected by the disease is modified due to the natural population increasing or to seasonal immigration. Denote by $R(t)$ the time function that gives the growth rate of the risk set. To take the population evolution effects into account in clustering studies, consider the transformation of the data set $(i, y_i)_{i=1,\ldots,N}$,

$$(i, \breve{y}_i) = (i, y_i \times R(x_i)) \quad \text{for } i = 1, \ldots, N. \quad (5)$$

Let us precisely determine what $R(t)$ should be. $R(t) = 1$ corresponds to a constant population. If the population size (n_0) increases regularly and doubles in 1 year, R should be an increasing function,

$$R(t) = 1 + \frac{1}{365} t \quad \text{for } \in [0, 365].$$

Thus, $R(t) \times n_0$ estimates the population size at time t.

For the Clusterville example, because the population doubles in the summer, the population risk rate function is

$$R(t) = 1 \times I_{[1,182) \cup [244,365]}(t) + 2 \times I_{[182,244)}(t) \quad \text{for } t \in [1, 365]. \quad (6)$$

The corresponding data set is $(i, \breve{y}_i) = (i, 10)$ for $i = 1, \ldots, 42$. The constant regression function $(y = 10)$ minimize the criterion and no clustering can be detected.

3. Applications

Cluster determination is important for detecting epidemics and for predicting patient numbers (e.g., in order to mobilize enough doctors in a hospital). The determination of seasonal clusters allows foreseeing an increase in the mortality number. Consequently, hospital managers can decide on the supplementary staff needed. Several options are proposed by the algorithm. Users can provide the minimal event number $C_i^{\min} := (k_i - l_i)$ that defines the cluster i. Observe that the algorithm does not necessarily need this number, but it decreases computational time. In the same way, the user can decide on a minimal distance between two successive clusters $D_i^{\min} := (l_i - k_{i+1})$.

Table 1 presents the algorithm used to determine one cluster and to decide about the uniformity hypothesis. The examples and simulations were implemented on a Ultra-Sparc Unix workstation using the S-Plus, Version 3.4, Release 1 software (MathSoft, 1996).

3.1 *Simulation Studies*

The first simulation intends to illustrate the criterion effect and to compare the proposed method with classical tests. The second example presents a case with multiple clusters.

3.1.1 *One-Cluster Detection.* The simulated data consists of one sample of N observations. The time events X are generated from a mixture of two uniforms: $N-10$ observations are sampled from U[0, 100] and 10 are sampled from U[35, 50]. One simulation is done with each value of N. On each of these data sets, we perform the test proposed by Tango (1984), the

Table 1
One-cluster C determination algorithm

Inputs: $X, R, C_{\min}$

$Y(i) \leftarrow X_{(i)} - X_{(i-1)}$ for $i \in \{2, \ldots, N\}$ and $Y(1) = X_{(1)}$
$\tilde{Y}(i) \leftarrow Y(i)R(i)$

for $k = 1$ to $N - C_{\min}$ do
  for $l = k + C_{\min}$ to N do
    $\hat{Y}(i) \leftarrow \bar{y}_{\bar{c}} \times \mathrm{I}_U(i) + \bar{y}_c \times \mathrm{I}_C(i)$
    $A(k,l) \leftarrow \sum_{i=1}^{N} (\tilde{Y}(i) - \hat{Y}(i))^2$
  end for
end for

$(\hat{k}, \hat{l}) \leftarrow \arg\min A(k,l)$

scan test, the variable scan test, and our method with different criterion choice. Results are given in Table 2. Note that, for these simulations, we used a window of size 15 for the scan test and six cases as the minimal size for the variable window, and we divided the interval by seven for Tango's index.

We use classical criteria, i.e., $BIC(i) = N \log((1/N)\|y_i - f(i)\|^2) + \log(N)d(i)$, where, in the penalty term $\log(N)d(i)$, $d(i)$ is the number of parameters of the model i. With no cluster, the regression function has one parameter estimated by $\bar{y}$; with one cluster, the regression function has four parameters estimated by $\bar{y}_{\bar{C}}$, $\bar{y}_C$, k, and l. With the AIC, the penalty term is $2d(i)$, and we also take $\log(\log(N))d(i)$, as usual in a multivariate context.

For the criterion, a large penalty term allows obtaining a conservative procedure, i.e., the simplest model with a few clusters is more significant with a large penalization ($\log(N)$) than with few ones (two or $\log(\log(N))$). According to our experience and because the results obtained are similar to those of the classical statistical tests, the AIC criterion seems well adapted. The BIC is more conservative, i.e., in the simulation (Table 2), the cluster presence is never significant. With the log-log penalty, the model with one cluster is always more significant than one with no clusters, even when its presence is not clear ($N = 50$).

3.1.2. *Multiple Clusters.* To illustrate the case with several distinct clusters, the simulated data consist of 100 samples of $N = 50$ observations. The time events X are generated as a mixture of uniform samples, i.e., 10 are sampled from U[30, 35], 10 from U[45, 50], 10 from U[60, 70], and 20 from U[0, 100].

Table 2
Simulation results

Model	$N = 25$	$N = 30$	$N = 35$	$N = 40$	$N = 50$
Tango's index	0.015	0.028	0.056	0.17	0.31
Scan test	0.042	0.065	0.09	0.14	0.29
Variable scan	0.03	0.055	0.09	0.08	0.12
$\alpha_{0,1}^{\mathrm{AIC}}$	0.004	0.025	0.03	0.048	0.087
$\alpha_{0,1}^{BIC}$	0.08	0.16	0.27	0.4	0.703
$\alpha_{0,1}^{\log\log}$	0.001	0.001	0.001	0.002	0.008

We perform the algorithm with minimal cluster sizes of $C_i^{\min} = 5$ events and a minimal distance between two clusters of two events (meaningful only for the model with multiple clusters). Thus, we obtain 100 values for $\alpha_{01}^{\mathrm{AIC}}$, $\alpha_{02}^{\mathrm{AIC}}$, $\alpha_{03}^{\mathrm{AIC}}$, $\alpha_{12}^{\mathrm{AIC}}$, $\alpha_{13}^{\mathrm{AIC}}$, and $\alpha_{23}^{\mathrm{AIC}}$. The boxplots of these values are shown in Figure 4.

According to these results, due to the few cases between two successive clusters, the model with only one cluster is selected ($\alpha_{01}^{\mathrm{AIC}} < 0.05$). Models with more than one cluster are always less significant than one with a single, large cluster. The percent of bootstrapped samples for which only one cluster is selected against the two- or the three-clusters model, $\alpha_{12}^{\mathrm{AIC}}$ or $\alpha_{13}^{\mathrm{AIC}}$, is not significant.

An example of the regression functions (Figure 3) shows how the method detects only one, two, or three clusters. Though the three clusters in [25, 35], [55, 65], and [75, 80] are simultaneously detected, the more significant model supposes a unique, large cluster.

3.2 *Knox Data Set*

The first real data set to illustrate the method consists of 35 cases of the birth defects esophageal atresia and tracheo-esophageal fistula observed in a hospital in Birmingham, United Kingdom, between 1950 and 1955. It was first published by Knox (1959) and subsequently analyzed by Weinstock (1981) using a scan statistic of fixed width and by Nagarwalla (1996) using a scan statistic with a variable window. The data can be found in the Appendix. It includes the number of days past January 1, 1950, on which each case was observed. The total time period of the study was 2191 days.

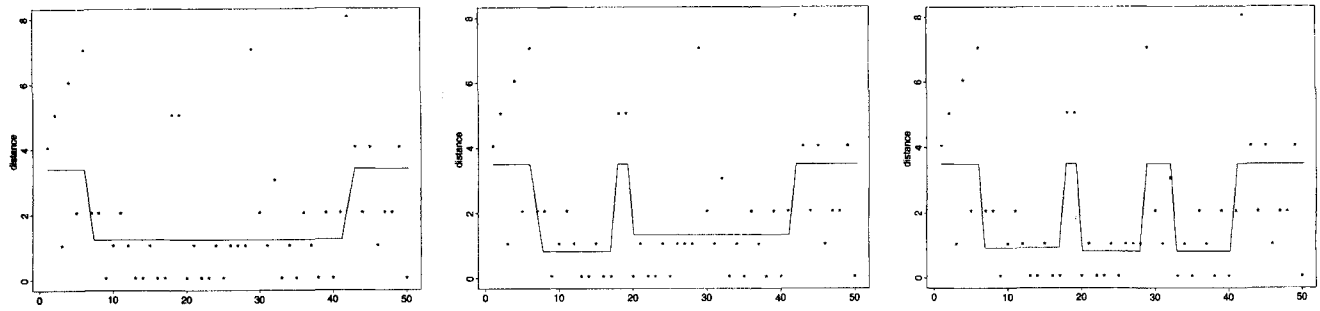

Figure 3. One sample and the estimation function with one, two, and three clusters.

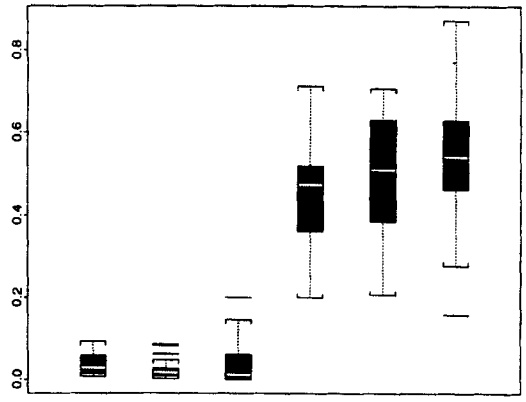

Figure 4. Boxplots for the α_{01}^{AIC}, α_{02}^{AIC}, α_{03}^{AIC}, α_{12}^{AIC}, α_{13}^{AIC}, and α_{23}^{AIC} values computed on 100 samples.

A comparison between the different tests used on this data set is summarized in Table 3. By using the AIC criterion, on each of the 1000 bootstrapped samples, the model with only one cluster is better than one with noncluster, $\alpha_{0,1}^{AIC} = 0$. The most likely cluster is the set of 15 cases beginning with the case on day 1233.

Moreover, the model with two clusters has a significant $\alpha_{0,2}^{AIC} = 0.001$ value compared with the nonclustering hypothesis. In fact, the two-cluster model detects the well-known cluster [1233, 1491] with 15 cases in 258 days and a second cluster [2049, 2174], with 7 cases in 125 days. According to these results, [2049, 2174] may be the beginning of a second cluster that is not significant compared with the existence of only one cluster $\alpha_{1,2}^{AIC} = 0.64$, perhaps because the study stops at day 2191.

Because the study time is large, this second cluster may be due to the natural increase of the population. In the next example, we illustrate our method by considering a variable population size.

3.3 *Hospital Hemoptysis Admission*

The method is used to detect clusters of minimum size $C_i^{min} = 6$ events. This number is chosen according to the Nice hospital pneumologic team, which estimated that more that six events in a short time can disturb the functioning of the hospital. Table 4 summarizes the results obtained for data presented in the Appendix.

The model with only one cluster minimizes the AIC criterion. The most likely cluster is the set of 14 cases from day 58 to day 87. Note that the model with two clusters provides a cluster in summer (day 187 to 201). This model can be selected against the constant model ($\alpha_{0,2}^{AIC} = 0.026$). Nice is in the south of France and had 355,000 inhabitants on January 1, 1995, and has a regular population increase of 0.72% per year. Moreover, there are 55,000 tourists in July and August (data provided by the tourist office of Nice). In this case, the corresponding $R(t)$ function is given by

$$R(t) = 1 + \frac{72t}{10{,}000 \times 365} + \frac{55{,}000}{355{,}000} \times I_{[182,244]}(t),$$

where $t \in [1, 365]$ is expressed in days. Results obtained on the transformed data set are presented in Table 4.

Table 3
Knox data set results

Model	p-Value	Cluster range
Tango's index	0.0009	
Scan test	0.017	
Variable scan	0.01	
$\alpha_{0,1}^{AIC}$	0	[1233, 1491]
$\alpha_{0,2}^{AIC}$	0.001	[1233, 1491] [2049, 2174]
$\alpha_{1,2}^{AIC}$	0.63	

Note that the model with one cluster is selected. The main difference between these results and those obtained without considering the evolution of the population at risk is as follows: The model with two clusters, the second one being in the summer, is not selected against the model with zero cluster. The fact that more events occur in [187, 201] is due to the presence of tourists.

The significant cluster is detected in winter (February 27–March 28). A study of the climatic factors in the Nice area shows that this period had very low temperatures (Berthier et al., 2000).

4. Discussion

The presented method allows one to detect several clusters during a study time. Based on a constant regression model, the data transformation provides an attractive visualization of the clusters. Our approach can easily be generalized for changes in the population at risk by modifying the data transformation according to the population growth rate.

The number of clusters and the number of cases are the important parameters in terms of computing time; the time period (number of days) has no effect on computations. For multiple clustering, because the algorithm is based on loops, it cannot, within a reasonable computational time, detect more than three clusters. For each of the data sets in Section 3.1.2, we need around 30 minutes (central processing unit [CPU] time) to obtain the results with the model supposing three clusters, 5 minutes for two clusters, and less than 30 seconds for detecting one cluster. Nevertheless, the introduction of minimal cluster size C_i^{min} and minimal event number between two successive clusters D_i^{min} allows reduction of the computation time.

The choice of the criterion is a much debated question. Müller (1992) presented a summary on this subject. In our context, with a small number of cases, we used the Akaike information criterion.

And last, note that the method has the potential to detect nondistinct clusters. Of course, this situation may not be of interest. Suppose two clusters are in the time periods $[a, c]$ and $[b, d]$, where $a < b < c < d$. In $[b, c]$, the two clusters are confounded. The piecewise regression function can be adjusted with the mean of the nonclustering period, two constants for the intervals $[a, b]$ and $[c, d]$, and a lower value in $[b, c]$.

A possible extension of this method to two-dimensional spatial clustering problems is to transform the data as in Section 2.1 to construct a distance between events. The Euclidean distance to the nearest neighbor may be an appropriate way to accomplish this. For space–time clustering,

Table 4
Hospital hemoptysis admission results

	Without transformation	Cluster range	With seasonal variation	Cluster range
$\alpha_{0,1}^{AIC}$	0.024	[58, 87]	0.02	[58, 87]
$\alpha_{0,2}^{AIC}$	0.026	[58, 126] [187, 201]	0.30	

the data points would be in three or higher dimensions, and one could also use the Euclidean distance.

Acknowledgements

The authors are grateful to Professor F. Boulay and Dr F. Berthier of DIM de Nice for providing data on the hemoptysis, to Marten Wegkamp for its corrections, and to the referees and the associate editor for helpful comments and suggestions.

Résumé

Dans cette article, nous présentons une méthode simple pour déterminer la présence d'un ou plusieurs agrégats d'événements temporaux au sein d'une population de taille variable. La méthode est basée sur un modèle de régression par morceaux qui après transformation des observations permet de tenir compte d'éventuelles modifications de la population à risque. Le choix entre les différents modèles et du nombre d'agrégats se fait par des procédures de rééchantillonnage. Nous avons appliqué cette approche à l'étude de la répartition de cas de maladies rares.

References

Akaike, H. (1974). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory,* B. N. Petrov and F. Csaki (eds), 267–281. Budapest: Akademiai Kiado Budapest.

Barton, D. E. and David, F. N. (1956). Some notes on ordered random intervals. *Journal of the Royal Statistical Society, Series B* **18,** 79–94.

Berthier, F., Boulay, F., Molinari, N., Daures, J. P., and Blaive, B. (2000). Rôle de la température sur l'existence de macro agrégats hivernaux d'hémoptysies. *Revue des Maladies Respiratoires* **17,** 1S125 (in French).

Cressie, N. A. (1977). The minimum of higher order gaps. *Australian Journal of Statistics* **19,** 132–143.

David, H. A. (1980). *Order Statistics, Wiley Series in Probability and Mathematical Statistics,* 2nd edition. New York: Wiley.

Ederer, F., Myers, E., and Mantel, N. (1964). A statistical problem in space and time: Do leukemia cases come in clusters? *Biometrics* **20,** 623–626.

Glaz, J. and Naus, J. (1983). Multiple clusters on the line. *Communications in Statistics—Theory and Methods* **12,** 1961–1986.

Huntington, R. J. and Naus, J. I. (1975). A simpler expression for kth nearest neighbor coincidence probabilities. *Annals of Probability* **3,** 894–896.

Hwang, F. K. (1977). A generalization of the Karlin–McGregor theorem on coincidence probabilities. *Annals of Probability* **5,** 814–817.

Knox, G. (1959). Secular pattern of congenital oesophageal atresia. *British Journal of Preventive Social Medicine* **13,** 222–226.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistical Theory and Methods* **26,** 1481–1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine* **14,** 799–810.

Larsen, R. J., Holmes, C. L., and Heath, C. W. (1973). A statistical test for measuring unimodal clustering: A description of the test and of its application to cases of acute leukemia in metropolitan Atlanta, Georgia. *Biometrics* **29,** 301–309.

MathSoft. (1996). *S-Plus for Unix Supplement,* Version 3.4. Seattle: Data Analysis Products Division.

McClure, D. E. (1976). *Extreme nonuniform spacings.* Report 44 in Pattern Analysis, Brown University, Providence, Rhode Island.

Müller, K. (1992). *Consistency properties of model selection procedures in multiple linear regression.* Technical Report 92-02, University of Montpellier, Montpellier, France.

Nagarwalla, N. (1996). A scan statistic with variable window. *Statistics in Medicine* **15,** 845–850.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association* **60,** 532–538.

Naus, J. I. (1966). Some probabilities, expectations and variances for the size of largest clusters and smallest intervals. *Journal of the American Statistical Association* **61,** 1191–1199.

Naus, J. I. (1982). Approximations for distributions of scan statistics. *Journal of the American Statistical Association* **77,** 177–183.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461–464.

Tango, T. (1984). The detection of disease clustering in time. *Biometrics* **40,** 15–26.

Wallenstein, S. (1980). A test for detection of clustering over time. *American Journal of Epidemiology* **111,** 367–372.

Weinstock, M. A. (1981). A generalized scan statistic test for the detection of clusters. *International Journal of Epidemiology* **10,** 289–293.

Whittemore, A. and Keller, J. B. (1986). A letter to the editor. On Tango's index of disease clustering in time. *Biometrics* **42,** 218.

Received October 1999. *Revised August* 2000.
Accepted September 2000.

APPENDIX

Knox Data Set

Cases of esophageal atresia and tracheoesophageal fistula over 2191 days from 1950 to 1955. The mean distance between adjacent cases $\bar{y}$ is 62.09 days, and the standard deviation is 86.48 days.

170 316 445 468 938 1034 1128 1233 1248 1249 1252 1259 1267 1305 1385 1388 1390 1446 1454 1458 1461 1491 1583 1699 1702 1787 1924 1974 2049 2051 2067 2075 2108 2151 2174.

Hospital Hemoptysis Admission Data Set

Days of hemoptysis admission at Nice University Hospital from January 1 to December 31, 1995, are as follows:

2 8 23 29 43 48 58 60 61 63 69 71 74 74 78 80 85 86 86 87 93 105 106 108 115 117 121 126 135 140 141 156 159 179 187 188 188 191 191 198 201 214 225 225 235 235 239 249 262 271 279 279 282 292 296 302 317 323 337 342 352 354.