# Temporal Clustering

**Hawkins, C.D. and D'Antuono, M.F.**
Western Australian Department of Agriculture and Food
Baron-Hay Court, South Perth, Western Australia 6151

## Abstract
When assessing the need for intervention in endemic disease management, it is useful to know if an occurrence of the disease of interest has exceeded the expected value over an appropriate time frame. The identification of a temporal cluster provides an indication for further investigation.

The Scan statistic has been used to identify temporal clustering, but has a number of shortcomings. In particular, the derivation is not intuitive and mathematically complex. The computational approximations that are available in various textbooks and papers may yield probability values less than zero, or greater than one, which leads one to question the approach.

Alternative methods are explored, including alternate probability models, and a Bayesian approach of using simulations to calculate the null probability distribution of no temporal clusters. Development of end user options such as a stand alone package is provided in the R Statistical System (freeware) called ScanoR.

## Introduction
A growing requirement of animal disease surveillance activities is the capacity to quickly and accurately determine whether an epidemic is occurring. The use of centralized disease recording and databases generates large quantities of data, with consequent limitations on human resources to search and analyse these data for clustering. Automated processes to evaluate clusters in time (temporal clustering) are needed to enhance the efficiency of surveillance. Detection of temporal clustering by the use of the Scan test offers some advance towards rapid detection of epidemics, and both spreadsheet methods (Carpenter and Ward 2003), and internet-based methods (Wallenstein 2005) have been made available. However, the methodology behind some of these computations is not intuitive, and may yield unrealistic values (i.e. outside the range of 0 to 1). Consequently, it is appropriate to reconsider the computation of a statistic that identifies temporal clustering in a more intuitive manner that yields probability values within an acceptable range.

## Materials and Methods
A temporal cluster in relation to endemic disease is identified on the basis of historical data, namely a number of events occurring over a retrospective period. The observational period of time is considered a window, and the essential element is whether a given number of observed events within one or more windows varies from a random scattering of events based on the historical data. In the approach taken here, random events were generated on the basis of historical data, and over multiple iterations the frequency of events occurring in nominated windows of time was calculated. The probability of these events was determined for the null distribution. Programs were written in R© (version 2.2.0 for Windows; R Development Core Team 2005), and a menu called ScanoR was provided using the R Commander library (Fox 2005).

We followed the approach by Wallenstein and Naus (2006) in assuming the null probability model may arise in a number of situations. At this stage the computer programs only look at a constant background rate of events and for the two scenarios of continuous and grouped data.

We denoted the "sc-Scan statistic" as the computation of a scan statistic S(w) for continuous data where we assumed that there were N=n events occurring on a time-line (0,T). We generated uniform samples from this time-interval and constructed an empirical distribution of the Pr(S(w)>=k) where k is the maximum number of events in a subinterval of width, w. It was

assumed that there are an infinite number of sliding windows on a time-line (0,T). However, our approach was to consider only the N=n points in a sample derived from this time-interval, and then we computed sliding windows on the ordered sample data values ($U_1$, … , $U_n$), using the windows [$U_i$, $U_i$+w) for i = 1, … , n. The left square bracket "[" indicates inclusion of the lower point; the open right parenthesis ")" indicates exclusion of the upper point, which is a typical convention in mathematical analysis. We also looked at the reverse chain of sliding windows [$U_j$ , $U_j$-w) for j = n, n-1, … , 1. By generating a large number of samples from the uniform distribution we derived an empirical distribution of Pr(S(w)>=k). Our interest was to see which value of "k" showed a small p-value, typically 0.05 or smaller.

Similarly, we denoted the "sg-Scan statistic" as the computation of a scan statistic S(w) for grouped data where we assumed that there were N=n events occurring over T disjoint time-intervals. We generated Poisson samples ($U_1$, … , $U_T$) such that the rate • = N/T. Then we summed the values $U_j$ for all j in each interval [i-1 , i-1+w), for i=1…T, and find the maximum of these sums. The intervals here are finite since there are only T intervals if size w=1, etc. For w=2, we computed sliding windows of size 2 from (2, 4 , …, T), and then (1, 3, …, T) in steps of 2. Similarly, for other values of increasing w. Naus and Wallenstein (2006) refer to their scan statistic for grouped data as the ratchet-scan statistic.

Again by generating a large number of samples from the Poisson distribution with rate "•" we derived an empirical distribution for Pr(S(w)>=k). Our interest was to see which value of "k" showed a small p-value, typically 0.05 or smaller.

A seed was used to start the random number generation and the seed is automatically inserted as the current time in (hours:minutes:seconds). The output echoes this so that repeated runs get the same answers.

The program was designated "ScanoR", to differentiate it from other scan statistics, and to acknowledge the R programming language.

Data from published Scan statistic results were reanalyzed for comparison.

## Results

The program presents the user with a dialogue box into which the historical data, the number of windows of interest, the desired number of iterations, and a random number seed can be entered (Figure 1). Default values are provided, but all may be varied as required. When the program has run, output appears as a sequence of possible numbers of events within the nominated window(s), with the probability of their occurrence listed below each number (Figure 2).

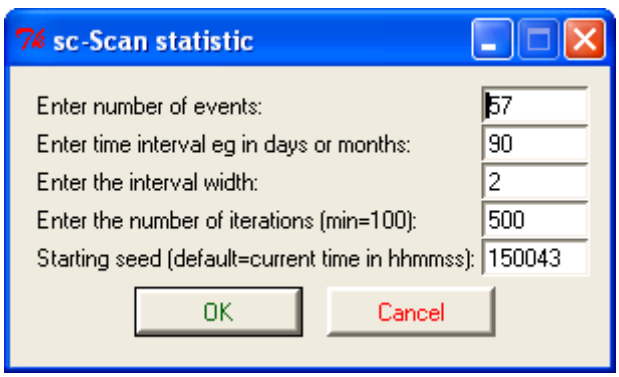


**Figure 1 Scan statistic dialogue box**

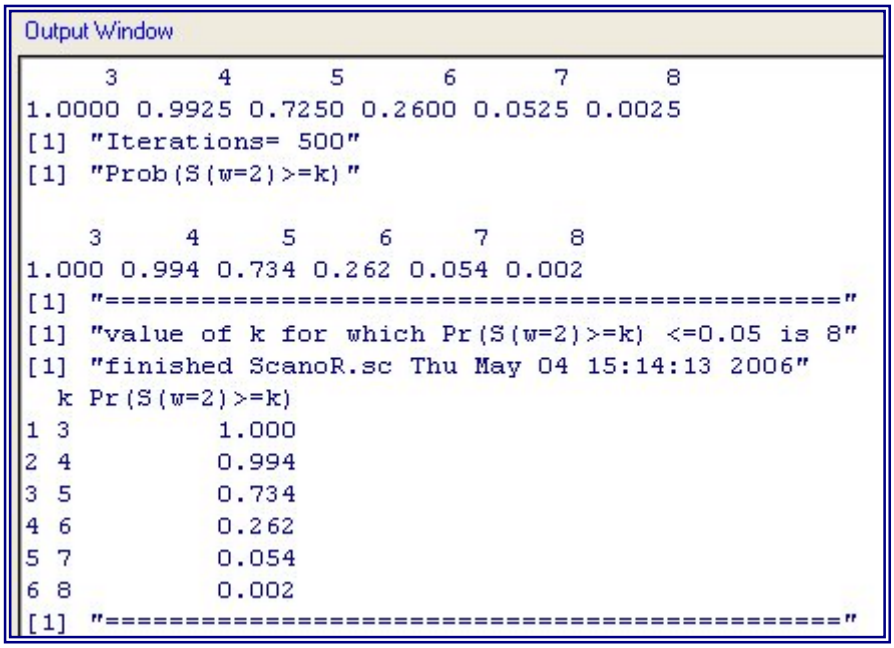


**Figure 2 sc-scan statistic output**

Data from Carpenter and Ward (2003), where 57 events were observed in a 90 month period, were reanalyzed using both the sc-Scan and the sg-Scan options. Results are given in Table 1, with the results using the method described by Carpenter and Ward (2003) included for comparison. At no stage does the methodology give values outside the range 0 – 1. Carpenter and Ward use the approximation by Wallenstein and Neff (1987) which assumes a continuous data framework, hence its agreement with the continuous data method.

**Table 1 Minimum number of events needed to identify an epidemic, given 57 events in a 90 month period, using a result of $\leq$ 0.05 as the discriminating criterion**

| Window width | Continuous Data method | Group Data method | Carpenter and Ward method |
|---|---|---|---|
| 1 | 6 | 5 | 6 |
| 2 | 8 | 7 | 8 |
| 3 | 9 | 8 | 9 |
| 4 | 10 | 9 | 10 |
| 5 | 11 | 10 | 11 |
| 6 | 12 | 12 | 12 |
| 7 | 13 | 13 | 13 |
| 8 | 14 | 14 | 14 |

The general agreement with an existing published approximation method is noted, although the grouped data scan is both consistent between the windows and gives smaller p-values for grouped data as described in Naus and Wallenstein (2006). With smaller numbers of iterations, and varying random number seeds, values may differ slightly from those given in Table 1. It is recommended that if a p-value is close to 0.05 then the program be re-run with more iterations (typically an extra

500 to 1000 iterations) to further refine the p-values. ScanoR is more informative than approximations since it provides more information about the various values of "k" found.

## Discussion

The ScanoR methodology at this stage looks at two scenarios and probability models for the null distribution in the time series of events, assuming a constant background rate of random events according to the null distribution. It is important to distinguish the appropriate model for the relevant sampling process. Two major concerns with existing methods or approximations of a scan statistic are, knowing what really is correct, and what assumptions are made about statistical distributions used in the approximation. As indicated previously, some existing approximations yield values outside the 0-1 range, and this casts some doubt on the appropriateness of underlying statistical assumptions. This is not the case with the ScanoR process, because of its iterative approach to determining probability values.

The ScanoR process and methodology offered here is internally consistent, transparent, flexible, simple to interface with data management systems, and may signal an intervention sooner than some other methods. Further refinements of the ScanoR process are envisaged, including automated linkages with animal health surveillance data to dynamically detect epidemics, as well as incorporating other model assumptions which may apply to the reference data.

## References


D'Antuono, M.F. (2006) ScanoR – Computations of Scan statistics using R©: A language and environment for statistical computing. Department of Agriculture and Food, Western Australia. http://www.agric.wa.gov.au (Email:mdantuono@agric.wa.gov.au)

Fox J. (2005) Rcmdr: A platform-independent basic-statistics GUI (graphical user interface) for R, based on the tcltk package. http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/

Naus, J. and Wallenstein, S. (2006) Temporal surveillance using scan statistics. Statistics in Medicine. Jan 30; 25(2):311-24.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Vose, David (2000) Risk Analysis. A quantitative guide. John Wiley and Sons, Chichester. 126-127

Wallenstein, S. (2005) Scan Statistic. http://c3.biomath.mssm.edu/wscan.html (subsequently removed)

Wallenstein S. and Naus J. (2003) Statistics for temporal surveillance of bioterrorism. In: Syndrome Surveillance: Reports from a National Conference, 2003. *Morbidity and Mortality Weekly Report* 2004;53 (Suppl), 74-78.

Wallenstein S. and Neff N (1987). An approximation for the distribution of the scan statistic. Statistics in Medicine 6: 197-207, cited in Ward MP and Carpenter TE *op. cit.*

Ward, M.P. and Carpenter, T.E. (2003) Methods for Determining Temporal Clusters in Surveillance and Survey Programs. *In:* Animal disease Surveillance and Survey Systems. Methods and Applications. Ed. Salman, M.D. Iowa State Press, Ames Iowa. Pp 87-99.