Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 23

5-1-2006

JMASM23: Cluster Analysis In Epidemiological Data (Matlab)

Andrés M. Alonso
Universidad Carlos III de Madrid, andres.alonso@uc3m.es

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 273-280

1538 – 9472/06/$95.00

JMASM23: Cluster Analysis In Epidemiological Data (Matlab)

Andrés M. Alonso
Department of Statistics
Universidad Carlos III de Madrid

Matlab functions for testing the existence of time, space and time-space clusters of disease occurrences are presented. The classical scan test, the Ederer, Myers and Mantel's test, the Ohno, Aoki and Aoki's test, and the Knox's test are considered.

Key words: Time cluster, space cluster, time-space cluster, epidemiology, Monte Carlo.

Introduction

The concept of groups or clusters of disease occurrences is enough well-known and intuitive. A cluster is defined as an unusual, real or perceived group of health events that are grouped in the time and/or in the space. Many triumphs in the control of infectious diseases have been the result of the epidemiological study of clusters of cases, for instance, the epidemic of cholera in London in the 1850s and the investigation of cases of pneumonia in Philadelphia in 1976 (legionary disease). The investigation of clusters of non-infectious diseases also has remarkable examples: dermatitis in people who use rings made with contaminated gold and vaginal carcinomas in women whose mothers who consumed diethylstilbestrol (see CDC, 1990).

The investigation of perceived clusters of health events requires the knowledge of some statistical instruments for determining if the observed group is real, taking into account the circumstances under study (the data type, the availability of comparison data, etc.). In this article, the aim is to describe some of the statistical techniques used to investigate clusters of health events and to provide Matlab routines that implement these techniques.

Andrés Alonso is a *Juan de La Cierva* Researcher at the Department of Statistics. His areas of research interest are statistical computing, resampling methods and biostatistics. E-mail: andres.alonso@uc3m.es.

Detection of Time Clusters

A time cluster is defined as a non-uniform distribution of the cases in the time interval for a given population under study. The objectives of these studies are:

1. To identify secular tendencies of the frequency of diseases in the populations.
2. To identify cyclical fluctuations in the occurrence of a disease.
3. To identify local epidemics of a disease.

Attention is focused on the methods related to the detection of local epidemics.

Scan test

The scan test is used to determine if the cases that appear in a geographic area are significantly near in time. The test statistics are the maximum number of events that happen in a time interval of fixed size t. This value is obtained by scanning in all the intervals of length t in the period under study. The critical values for this test are provided in the tables calculated by Naus (1965, 1966) and Wallenstein (1980).

It is assumed that T is the complete observational interval and t is the duration time of one epidemic. Let be $r = t/T$, N the number of cases that happened in time T, and $p = \Pr(n, N, r)$ is the probability that a maximum number of cases in any interval of length t exceeds or is equal to n. This probability is calculated under the hypothesis that the N events are uniformly distributed in the interval T. The problem consists of estimating p. Wallenstein (1980) proposed the following algorithm: If the observed interval is a multiple of 12, 24, 36, 48 or 60 months, and if the duration of the epidemic is a multiple from 2 to 4 or 6 months, many quotients $r=t/T$ can be reduced to the fraction $1/L$ with $L = 4, 6, 8, 12, 15$ or the 24. If N is greater than 10 and smaller than 100, then tables in Wallenstein (1980) give the critical values of the distribution of n.

Example 1: The following table shows the number of cases of trisomia and spontaneous abortion in the city of New York between July/1975 and June/1977 (see Bailar et al., 1970).

```
function p = ProbabilityOfScanTest(n, N, t, T, B)

% Inputs:
% -------
% n : Maximum number of cases observed in t periods.
% N : Number of cases observed in T periods.
% t : Epidemic duration time.
% T : Total observation time.
% B : Number of replications.
%
% Output:
% -------
% p : Probability of having a value bigger or equal to n.

% Cases are B independents replicas of a uniform distribution
% of N cases in T periods.
Cases = zeros(T, B);
for b = 1:B
   X = rand(N, 1);
   for ii = 1:N
      for tt = 1:T
         if ((tt-1)/T < X(ii, 1) & X(ii, 1) < tt/T)
            Cases(tt, b) = Cases(tt, b) + 1;
         end
      end
   end
end

% Calculating the scan statistics using the B generates replicas
% stored in variable Cases.
ScanStatistics = zeros(B, 1);
for b = 1:B
   for tt = 1:T-t+1
      if (ScanStatistics(b, 1) < sum(Cases(tt:tt+t-1, b)))
         ScanStatistics(b, 1) = sum(Cases(tt:tt+t-1, b));
      end
   end
end

% Estimating the probability of having a scan statistics bigger
% or equal to the observed value, n.
p = sum(ScanStatistics >= n)/B;
```

Figure 1. Matlab Function `p`

Month / Year	Cases
07/1975 – 12/1975	0, 4, 1, 2, 1, 3
01/1976 – 06/1976	1, 3, 2, 2, 3, 4
07/1976 – 12/1976	1, 1, 1, 2, 4, 7
01/1976 – 06/1976	7, 2, 2, 6, 1, 2

Therefore, N = 62, T = 24 months and the epidemic duration is fixed to t=2 months. Then n=14 and Pr(14,62,1/12) can be calculated. The Matlab function in Figure 1 obtain the probability p = Pr(n, N, r) by a Monte Carlo simulation procedure. The results of the above function for the data in Example 1 is Pr(14,62,2,24)= 0.0113. It supports the conclusion of a time cluster.

Test of Ederer, Myers and Mantel

The period under study is divided in k disjoints intervals. Under the null hypothesis of no grouping, the n cases will have to be distributed uniformly in the k intervals. The test statistics, m, is the maximum number of cases in an interval. Mantel et al. (1976) calculated tables for the expectation and variance of m under the null hypothesis of no group and for selected values of k and n. In the following table, the approximated estimators of E (m) and Var(m) are shown when the number of cases is greater than 100 (see Mantel et al., 1976).

Number of intervals, k	$E(m)$	$Var(m)$
2	$n/2 + 0.3989 * n^{½}$	$0.09084 * n$
3	$n/3 + 0.4886 * n^{½}$	$0.07538 * n$
4	$n/4 + 0.5147 * n^{½}$	$0.06043 * n$
5	$n/5 + 0.5201 * n^{½}$	$0.04951 * n$

Example 2: Assume that the number of children with congenital malformations born in the same year is as follows: 1st trimester: 100 cases, 2nd trimester: 50 cases, 3rd trimester: 50 cases and 4th trimester: 70 cases. If k=4 and n=270, then one can use the estimators of the previous table: *E(m)*= 270/4+0.5147*√270 ≈ 75.95 and *Var(m)* =0.06043*270 ≈ 16.32. The following statistic is calculated,

$$\chi = \frac{(m - E(m))^2}{Var(m)} = \frac{(100 - 75.95)^2}{16.32} \approx 35.44\,,$$

and it may be concluded that it exists a time cluster.

```
function [E, V] = EdererMyersMantelTest(m, n, k, B)

% Inputs:
% -------
% m : Maximum number of cases observed in one interval.
% n : Number of cases observed in the period under study.
% k : Number of intervals.
% B : Number of replications.
%
% Output:
% -------
% E : Expected value of m.
% V : Variance of m.

% Cases are B independents replicas of a uniform distribution
% of n cases in k intervals.
Cases = zeros(k, B);
for b = 1:B
   X = rand(n, 1);
   for ii = 1:n
      for tt = 1:k
         if ((tt-1)/k < X(ii, 1) & X(ii, 1) < tt/k)
            Cases(tt, b) = Cases(tt, b) + 1;
         end
      end
   end
end

% Calculating the maximum m using the B generated replicas
% stored in variable Cases.
mStatistics = max(Cases);

% Estimating the mean and the variance of m.
E = mean(mStatistics);
V = var(mStatistics);
```

Figure 2. Matlab function `[E, V]`

The Matlab function in Figure 2 obtains the estimators of $E(m)$ and $Var(m)$ by a Monte Carlo simulation procedure. The results of this function for the data in Example 2 is $E(m) = 76.07$ and $Var(m) = 17.52$.

Detection of Space Clusters

A space cluster is defined as a non-uniform distribution of the cases in the area under study relative to the distribution of the population under study. The presence of clusters suggests a possible environmental etiology. The simplest analysis of space cluster is the comparison of the incidence or the prevalence of a particular disease in different geopolitical areas.

Test of Ohno, Aoki and Aoki

The test proposed by Ohno et al. (1979) determines if the obtained geographic pattern is different from the expected geographic pattern under the assumption of a uniform random distribution of the cases in the area under study. The procedure is as follows:

1. Define $k > 2$ disjoint categories of the incidence rates.

2. Identify the adjacent geographic areas in a map of the area under study.
3. Count the number of concordant area pairs.
4. Calculate the expected number of concordant adjacent pairs for each category: Let be N the number of areas and N_i the number of areas in the i-th category, then the number of concordant pairs in category i is $N_i(N_i-1)/2$. Let A be the number of adjacent pairs of regions, then the expected number of adjacent pairs with the i-th category is

$$E(C_i) = \frac{A}{N(N-1)} N_i(N_i - 1) .$$

5. Calculate the expected number of concordant adjacent pairs:

$$E(C) = \sum_{i=1}^{k} E(C_i) .$$

Finally a χ^2 test statistics, $\chi^2 = \frac{(C - E(C))^2}{E(C)}$, is calculated.

Example 3: The mortality rates of vesicle and esophagus cancer in Japan (1967-71) is categorized according to the following criterion:

Category 1. Rate≥140 by 10000 inhabitants.
Category 2. 120≤Rate≤139.9 by 10000 inhabitants.
Category 3. 80≤Rate≤119.9 by 10000 inhabitants.
Category 4. 60≤Rate≤79.9 by 10000 inhabitants.
Category 5. Rate≤60 by inhabitants.

In 1970, Japan had N = 1,123 cities and towns, without counting the prefecture of Okinawa, with A=2840 adjacent pairs of regions. The number of regions by category was: N_1 = 293, N_2 = 78, N_3 = 256, N_4 = 116 and N_5 = 380. In the following table, the calculation required for Ohno, Aoki and Aoki's test is presented.

Concordant pairs	Observed, C_i	Expected, $E(C_i)$	χ^2
(1,1)	201	192.84	0.35
(2,2)	17	13.54	0.89
(3,3)	170	147.14	3.55
(4,4)	25	30.07	0.85
(5,5)	315	324.61	0.28
Total	728	708.20	0.55

Finally, χ^2=0.55 and it is concluded that evidence does not exist for the geographic association of the vesicle and esophagus cancer in men for these years in Japan. The following Matlab function obtain the value of Ohno, Aoki and Aoki's test statistics given N, A, C and the N_i.

```
function OAAtest = OhnoAokiAokiTest(N, A, Ni, C)

% Inputs:
% -------
% N : Total number of regions.
% A : Number of adjacent regions.
% Ni : Number of regions in the ith category (k x 1 vector).
% C : Observed number of concordant adjacent regions.
%
% Output:
% -------
% OOAAtest : Ohno, Aoki and Aoki test statistics.

% Numbers of categories.
k = length(Ni);

% Expected number of adjacent regions in the ith category.
ECi = A*Ni.*(Ni-1)/(N*(N-1));

% Expected number of concordant adjacent regions.
EC = sum(ECi);

% Ohno, Aoki and Aoki test statistics.
OAAtest = (C-EC)^2/EC;
```

Figure 3. Matlab Function `OAAtest`

Detection of Space-Time Clusters

A space-time cluster is defined as a non-uniform distribution of the cases in space and time, simultaneously. In general, the test of space-time cluster of health events needs a more a more sophisticated elaboration because one needs to prove that if the cases are associated in space they are also significantly near in the time, and vice versa (see, e.g., Kleinbaum et al., 1982).

Test of Knox

The test proposed by Knox (1964) is used to determine if there exists a significant interaction between the sites and the moments of appearance of the disease. It divides the dimensions in space-time into two parts, for which the critical distance in space, E, and the critical distance in time, T, must be defined. In a contingency table, each pair of cases is classified in one of the following categories: (i) near only in space, (ii) near only in time, (iii) near in space-time, and (iv) distant both in space and in time. The procedure is as follows:

1. Let be n the number of cases. For each case, one knows its position in the space and in the time, then there are $N = n(n-1)/2$ possible pairs of cases.
2. Determine the distances in space, e, and in time, t, for each pair of cases.
3. Classify the N pairs according to the following criterion:

   (a) A pair is near in space if $e<E$.

   (b) A pair is near in time if $t<T$.

   (c) A pair is near in space-time if it fulfills (a) and (b), simultaneously.

   (d) When a pair satisfies neither (a) nor (b), then we say that it is not near nor in space nor in time.

4. Construct the following table:

	Space		
Time	Near	Non-Near	Total
Near	X	N_t - X	N_t
Non-Near	N_e - X	N - N_t - N_e+X	N - N_t
Total	N_e	N - N_e	N

where N_e is the number of pairs near in the space, N_t the near ones in the time, and X the near pairs in space-time.

5. The test statistic is the observed number of pairs near in space-time, X. In Knox (1964) it is assumed that X distributes as a Poisson, therefore,

$$p = \Pr(X \geq x) = \sum_{i=x}^{N} \frac{e^{-\lambda}\lambda^{i}}{i!},$$

where $\lambda = N_e N_t / N$.

Example 4: The following table shows the results of the method of Knox for 5 cases of meningococcal disease in a territory given in a period of one year, it takes like critical distance in space 500 meters and in time 5 days.

	Space		
Time	Near	Non-Near	Total
Near	X=4	0	N_t= 4
Non-Near	1	5	6
Total	N_e= 5	5	N=10

Therefore, λ=5*4/10 = 2 and Pr(X≥4) = 0.142. The Matlab function in Figure 4 obtains the value of above p-value given X, N_e, N_t and N.

```
function pKtest = KnoxTest(X, Ne, Nt, N)

% Inputs:
% -------
% X : Number of pairs near in space-time.
% Ns : Number of pairs near in space.
% Nt : Number of pairs near in time.
% N : Total number of pairs.
%
% Output:
% -------
% pKtest : Pvalue of Knox test statistics.

% Parameter of the Poisson distribution.
lambda = Ne*Nt/N;

% p-value of Knox test statistics.
pKtest = 0;
for i = X:N
   pKtest = pKtest + exp(-lambda)*lambda^i/factorial(i);
end
```

Figure 4. Matlab Function `pKtest`

References

Bailar, J. C., Eisenberg, H. & Mantel, N. (1970). Time between pairs of leukemia cases. *Cancer*, *25*, 1301-1303.

CDC: Center of Disease Control (1990). Guidelines for investigating clusters of health events, *MMWR*, *39*.

Kleinbaum, D. G., Kupper, L. L. & Morgenstern, H. (1982). *Epidemiologic research, principles and quantitative methods*. New York, N.Y.: Van Nostrand Reinhol Company.

Knox, G. (1964). Detection of space-time interactions. *Applied Statistics*, *13*, 25-30.

Mantel, N., Kryscio, R. J. & Myers, M. H. (1976). Tables and formulas for extended use of Ederer-Myers-Mantel disease-clustering procedure. *American Journal of Epidemiology*, *104*, 576-588.

Naus, J. (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 the largest clusters and smallest intervals. *Journal of the American Statistical Association*, *61*, 1191-1199.

Ohno, Y., Aoki, K. & Aoki, N. (1979). A test of significance for geographic clustering of disease. *International Journal of Epidemiology*, *8*, 273-281.

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