Statistical techniques for local cluster detection

Alexandru Amărioarei

National Institute of Research and Development for Biological Sciences

2^{nd} BIS Workshop: bioinformatic and statistical tools for data analysis

Friday 10 June, 2016

Research Platform for Biology and Systemic Ecology
Conference Hall, Spl. Independentei no 91-95, Bucharest, Romania

Outline

Outline

Education

Alexandru Amărioarei

University of Science and Technologies, Lille, France 2014

- Ph.D. Thesis: *Approximations for Multidimensional Discrete Scan Statistics*
- Advisor: Prof. Cristian Preda
- Member of MODAL Team - Models for Data Analysis and Learning Team (INRIA Lille)

University of Bucharest, Bucharest, Romania 2008–2010

- Master Thesis: *Markov chains with applications in biology* (in romanian)
- Advisor: Acad. Ioan Cuculescu

University of Bucharest, Bucharest, Romania 2004–2008

- Bachelor Degree (Mathematics)

Outline

Research Interests

Research Topics

- Scan statistics: methods and applications
- Distribution of runs and patterns
- Simulation techniques based on Monte Carlo methods
- Scientific computing
- Spatial Data Analysis
- Concentration Innequalities

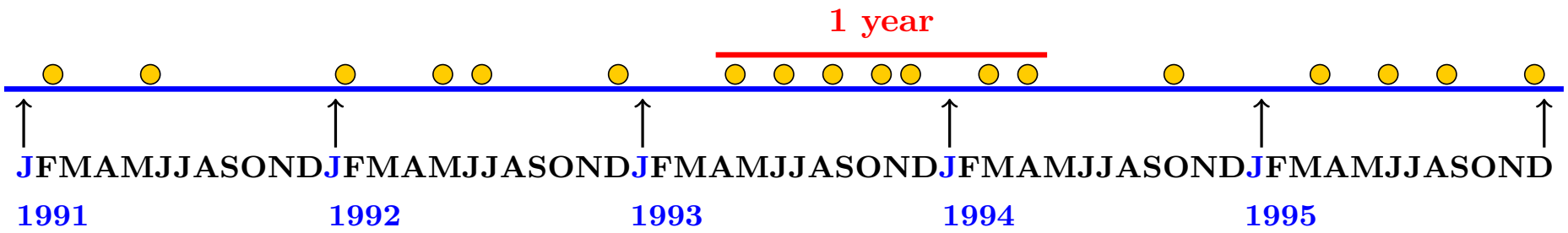
Languages

- Matlab
- R
- SAS
- Mathematica
- Maple

Outline

A first example

Example from Epidemiology



Observation of disease cases over time:
N = 19 cases over a period of $T = 5$ years

Observation

The epidemiologist notes a **one year period** (from April 93 - through April 94) with **8** cases: 42%!

Question

Given 19 cases over 5 years, how unusual is it to have a 1 year period containing as many as 8 cases?

The answer: A First approach

A First approach:

$X =$ the number of cases falling in [April 93, April 94]

$X \sim Bin(19, 0.2)$

$$\mathbb{P}(X \geq 8) = 0.023$$

Conclusion: **an atypical situation !**

But: it is **not** the answer to our question: the one year period is **not fixed** but identified after the scanning process !

The answer: Correct approach

The scan statistics:

S = the maximum number of cases over **any continuous** one year period in $[0, T]$

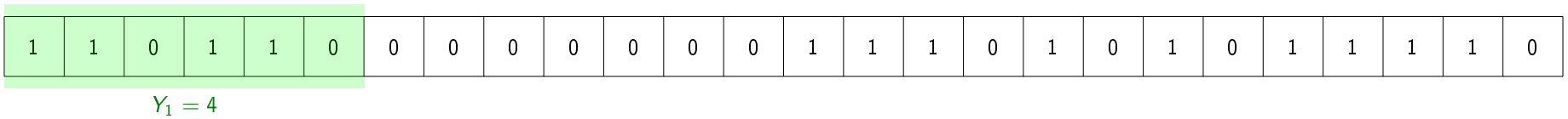
Thus,

$$\mathbb{P}(S \geq 8) = 0.379$$

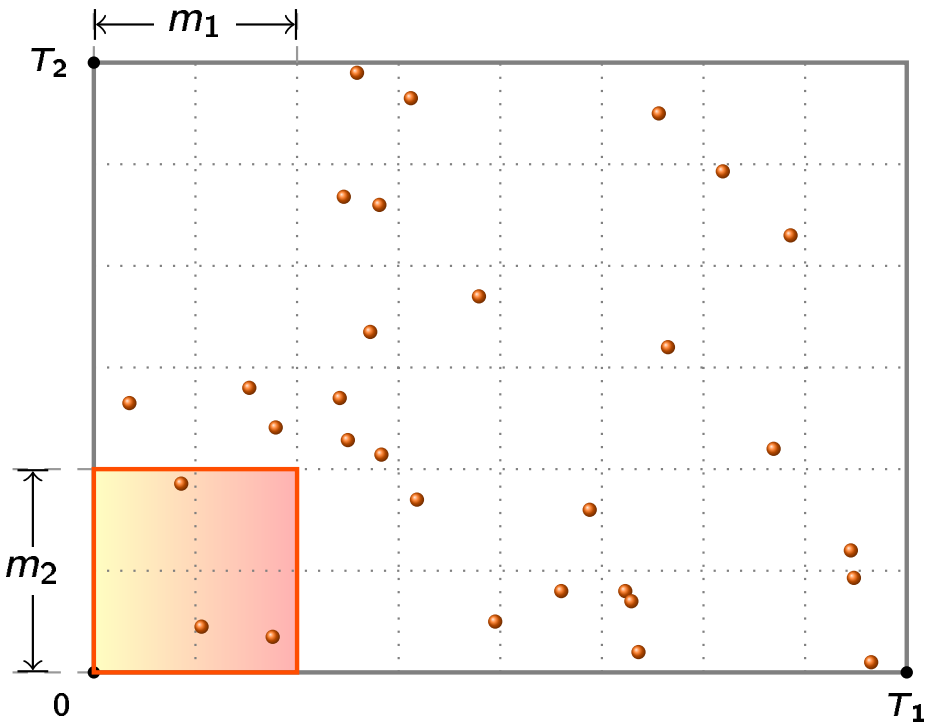
gives the answer to the epidemiologist question.

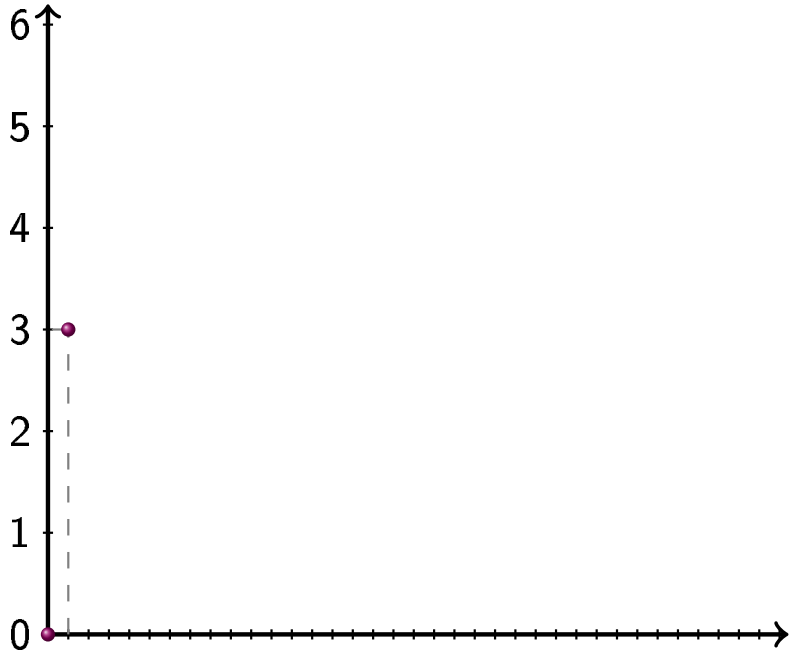
Conclusion: **no unusual situation !**

Example



Animation for 2 dimensional scan statistics





Number of points: 3

Outline

Detection of Crohn's disease clusters using spatial scan statistics

Problem and Data

Crohn's disease(CD) - an inflammatory disease of the intestines, which has no known pharmaceutical or surgical cure

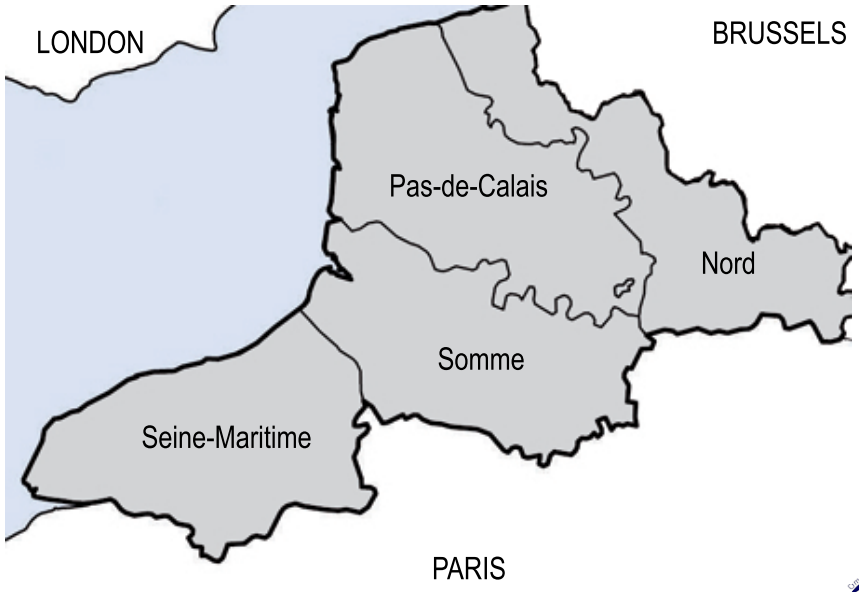
- genetic factors
- environmental risk factors

Goal

Detect and highlight significant atypical clusters of CD in terms of incidence

Data

- Study region: North of France
- Population: 5 790 526
- Period: 1990–2006
- Sub-division (cantons): 273 (small French administrative area with population between 1 500 and 212 000)
- Per canton: stratified population (gender and age group)
- Cases of Crohn's disease: 6 472



Methods

Standardized Incidence Ratios (SIR) [Declercq et al., 2010, Besag and York, 1991]

- detect global spatial heterogeneity
- unable to detect unusual local clusters of CD
- unable to test their significance
- cannot take into account the time component

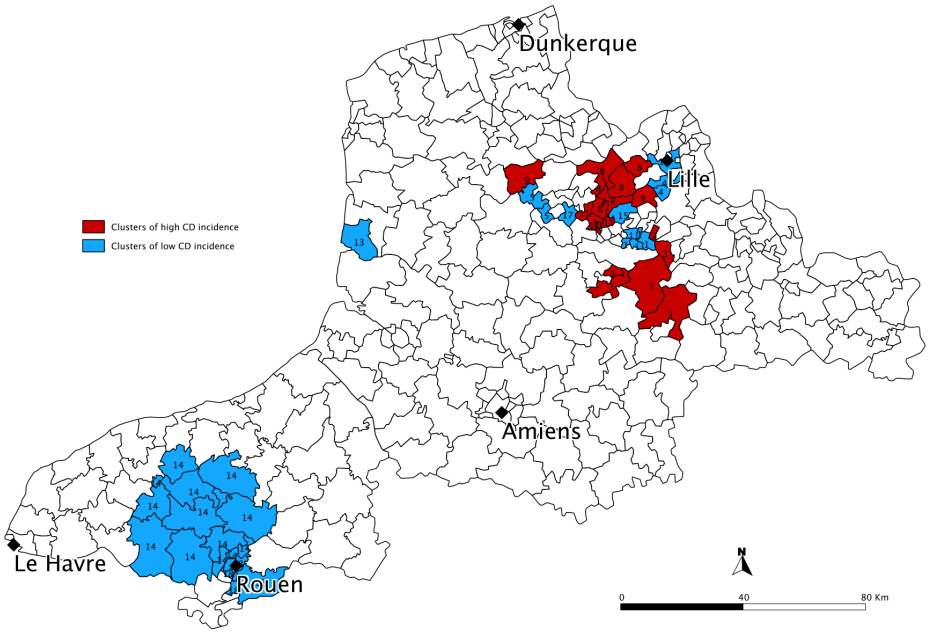
▸ SIR

Spatial and space-time scan statistics[Kulldorff, 1997, Kulldorff, 2006]

- detect local clusters without pre-selection bias
- detection of time-constant clusters
- detection of time-varying clusters
- able to test their significance

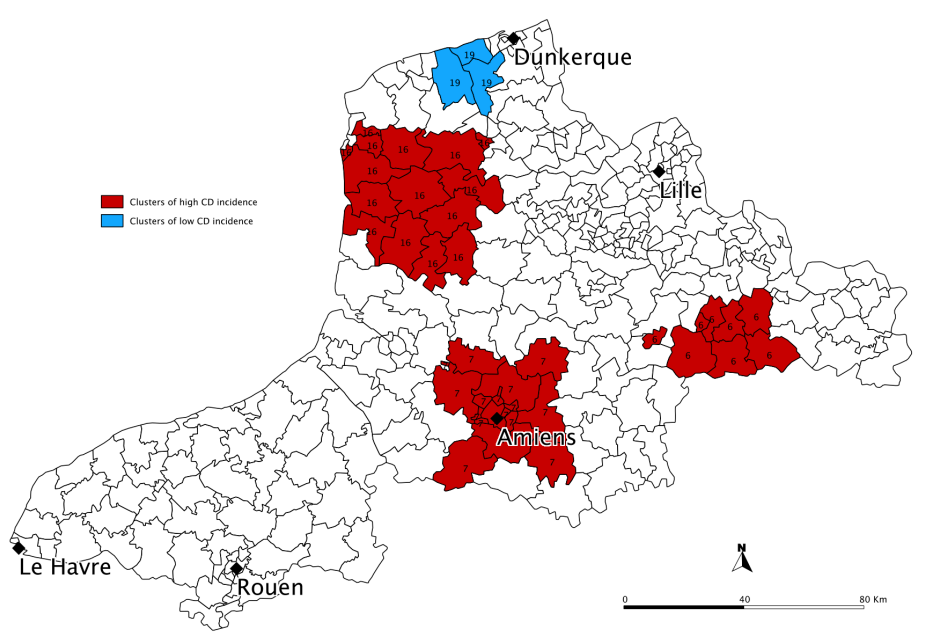
▸ Spatial scan statistics

Results



Time-constant clusters

- 14 significant clusters detected
- 5 clusters with high incidence (total: 726 cases)
- 9 clusters with low incidence (total: 521 cases)



Time-varying clusters

- 4 significant clusters detected
- 3 clusters with high incidence (779 cases within a period from 9 to 12 years)
- 1 clusters with low incidence (4 cases over 7 years period)

thank you!

A mathematical model for matching in two aligned sequences

Matching in two aligned sequences

Let $\{Y_1, Y_2, \ldots, Y_{T_1}\}$ and $\{Z_1, Z_2, \ldots, Z_{T_1}\}$ be two i.i.d. sequences of r.v.'s over the four-letter alphabet $\mathcal{A} = \{A, C, G, T\}$.
Define for $1 \leq i \leq T_1$, the score r.v.'s

$$X_i = \begin{cases} 1, & \text{if } Y_i = Z_i \\ 0, & \text{otherwise} \end{cases}, \quad X_i \sim \mathcal{B}(p), \quad p = \mathbb{P}(Y_i = Z_i)$$

Let V_c denote the length of the longest matching subsequence allowing at most c mismatches.

Example ($T_1 = 26$, $p = 0.25$, $c = 1$)

Y:	A	A	A	C	C	G	G	G	C	A	C	T	A	C	T	T	T	G	A	G	A	C	G	T	G	A
Z:	A	A	T	C	C	C	C	C	G	T	G	C	C	C	T	T	A	G	C	G	G	C	G	T	G	G
X:	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	1	1	1	1	0

- $c = 0$: length of the longest success run L_{T_1} ([Bateman, 1948])
- $c \in \{1, 2\}$: almost perfect run ([Han and Hirano, 2003], [Bersimis et al., 2012])

Matching in two aligned sequences

Let $\{Y_1, Y_2, \ldots, Y_{T_1}\}$ and $\{Z_1, Z_2, \ldots, Z_{T_1}\}$ be two i.i.d. sequences of r.v.'s over the four-letter alphabet $\mathcal{A} = \{A, C, G, T\}$.
Define for $1 \leq i \leq T_1$, the score r.v.'s

$$X_i = \begin{cases} 1, & \text{if } Y_i = Z_i \\ 0, & \text{otherwise} \end{cases}, \quad X_i \sim \mathcal{B}(p), \quad p = \mathbb{P}(Y_i = Z_i)$$

Let V_c denote the length of the longest matching subsequence allowing at most c mismatches.

Example ($T_1 = 26$, $p = 0.25$, $c = 1$)

Y:	A	A	A	C	C	G	G	G	C	A	C	T	A	C	T	T	T	G	A	G	A	C	G	T	G	A
Z:	A	A	T	C	C	C	C	C	G	T	G	C	C	C	T	T	A	G	C	G	G	C	G	T	G	G
X:	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	1	1	1	1	0

- $c = 0$: length of the longest success run L_{T_1} ([Bateman, 1948])
- $c \in \{1, 2\}$: almost perfect run ([Han and Hirano, 2003], [Bersimis et al., 2012])


Bateman, G. (1948).
On the power function of the longest run as a test for randomness in a sequence of alternatives.
Biometrika, 35:97–112.


Bersimis, S., Koutras, M. V., and Papadopoulos, G. (2012).
Waiting time for an almost perfect run and applications in statistical process control.
Methodol Comput Appl Probab.


Besag, J. and York, J. (1991).
Bayesian image restoration with two applications in spatial statistics.
Annals of the Institute of Statistical Mathematics, 43:1–21.


Breslow, N. and Day, N. (1980).
Statistical methods in cancer research.
The analysis of case-control studies: Distributed for IARC by WHO.


Declercq, C., Gower-Rousseau, C., Vernier-Massouille, G., Salleron, J., Balde, M., Poirier, G., Lerebours, E., Dupas, J. L., Merle, V., Marti, R., Duhamel, A., Cortot, A., Salomez, J. L., and Colombel, J. F. (2010).
Mapping of inflammatory bowel disease in northern france: spatial variations and relation to affluence.

Inflamm Bowel Dis, 16(5):807–12.

Han, Q. and Hirano, K. (2003).

Waiting time problem for an almost perfect match.

Stat. and Prob. Letters, 65:39–49.

Kulldorff, M. (1997).

A spatial scan statistic.

Communications in Statistics - Theory and Methods, 26(6):1481–1496.

Kulldorff, M. (2006).

Tests of spatial randomness adjusted for an inhomogeneity.

Journal of the American Statistical Association, 101(475):1289–1305.

Samuels, S., Beaumont, J., and Breslow, N. (1991).

Power and detectable risk of seven tests for standardized mortality ratios.

American journal of epidemiology.

Standard Incidence Ratio

SIR: is defined as the ratio between O_i, the number of observed cases in region (canton) i over the studied period and the expected number of cases E_i under the incidence rate hypothesis adjusted by sex and age group over the reference population.

- n_{ijk} - population for the i^{th} region (canton) with age class j and sex k
- λ_{jk} - incidence ratio for the age class j and sex k

$$E_i = \sum_j \sum_k \lambda_{jk} n_{ijk}$$

The *standardized incidence ratio* relative to region i:

$$SIR_i = \frac{O_i}{E_i}$$

with $\mathbb{E}[SIR_i] = \theta_i$ and $\mathbb{V}[SIR_i] = \frac{\theta_i}{E_i}$ estimated by $\frac{O_i}{E_i^2}$

Return

Standard Incidence Ratio: Interpretation

Interpretation

- $SIR_i = 1$: the incidence in the region (canton) i is not different than the expected one in the reference population (no risk)
- $SIR_i > 1$: the incidence in the region (canton) i is higher than the expected one in the reference population
- $SIR_i < 1$: the incidence in the region (canton) i is lower than the expected one in the reference population

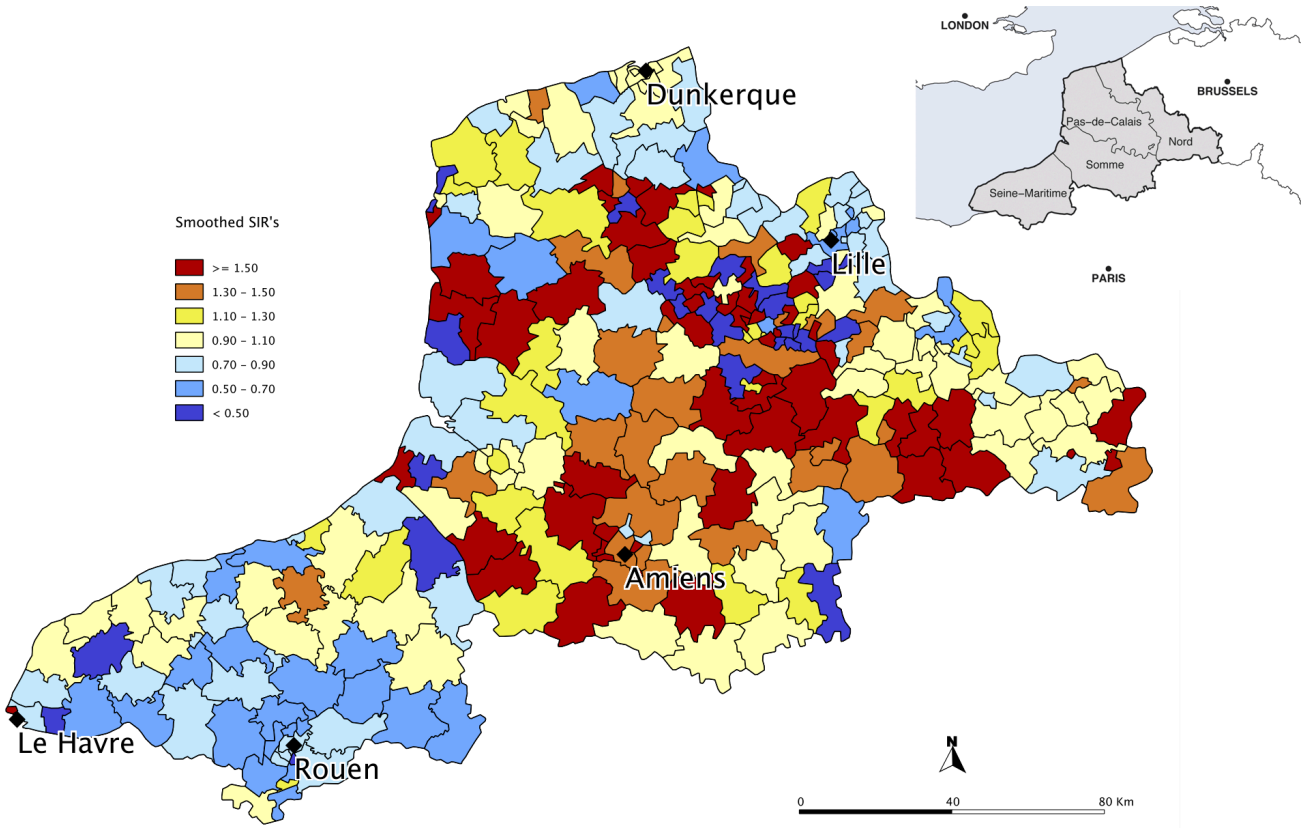
Statistical Test

- H_0: $SIR = 1$
- H_1: $SIR \neq 1$

Test statistics [Breslow and Day, 1980] and [Samuels et al., 1991].

Return

SIR: Crohn's disease example



Spatial scan statistics

Assumption

The number of CD cases in each canton is Poisson distributed

- **The null hypothesis**: the risk of being affected by CD is constant throughout all cantons
- **The alternative hypothesis**: there is at least one region for which the underlying risk is higher inside the region as compared to outside

Description:

- circular window of flexible size (varying form 0 up to a maximum radius so that the window never contains more than 50% of the population-at-risk)
- uses as center of the window the centroid of the cantons
- for each circle, the likelihood to observe the number of CD cases within and outside is computed and the circle, which maximizes the likelihood, is defined as the *most likely cluster* (MLC)

Return

Spatial scan statistics: likelihood

Under a Poisson model, the likelihood of a zone Z is given by:

$$L(Z) = \frac{e^{-n_G}}{n_G!} \left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(\frac{n_G - n_Z}{\mu(G) - \mu(Z)}\right)^{n_G - n_Z} \prod_{i=1}^{n} \mu(d_i)$$

where $d_1, d_2, \ldots, d_n$ are the sites locations (centroid), $\mu(d_i)$ is the population at risk in the location d_i and n_Z, $\mu(Z)$, n_G, $\mu(G)$ are the number of CD cases and the population at risk inside the circular zone Z and in the whole region G.
The test statistic used is

$$\nu = \max_Z \frac{L(Z)}{L_0}$$

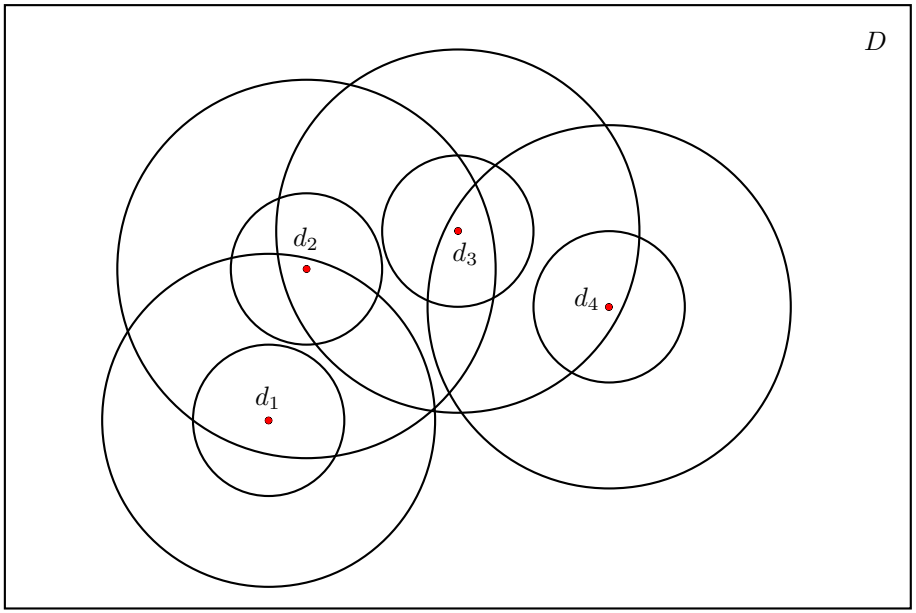
where the likelihood under the null hypothesis is

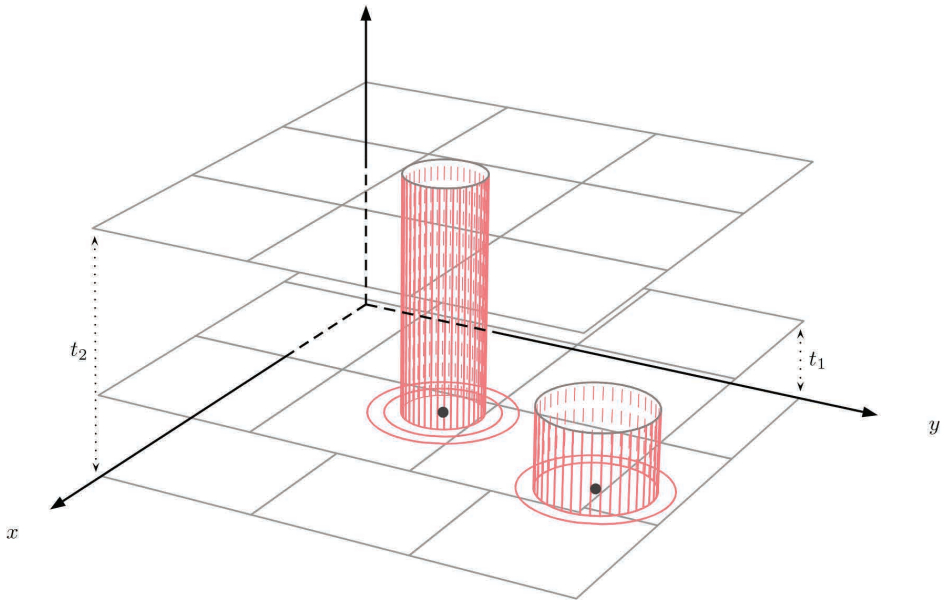
$$L_0 = \frac{e^{-n_G}}{n_G!} \left(\frac{n_G}{\mu(G)}\right)^{n_G} \prod_{i=1}^{n} \mu(d_i)$$

The p-value, $\mathbb{P}(\nu > \nu_{obs})$, associated to the MLC is obtained based on Monte-Carlo random replications under the null hypothesis.

Return

Spatial scan statistics: illustration





Return