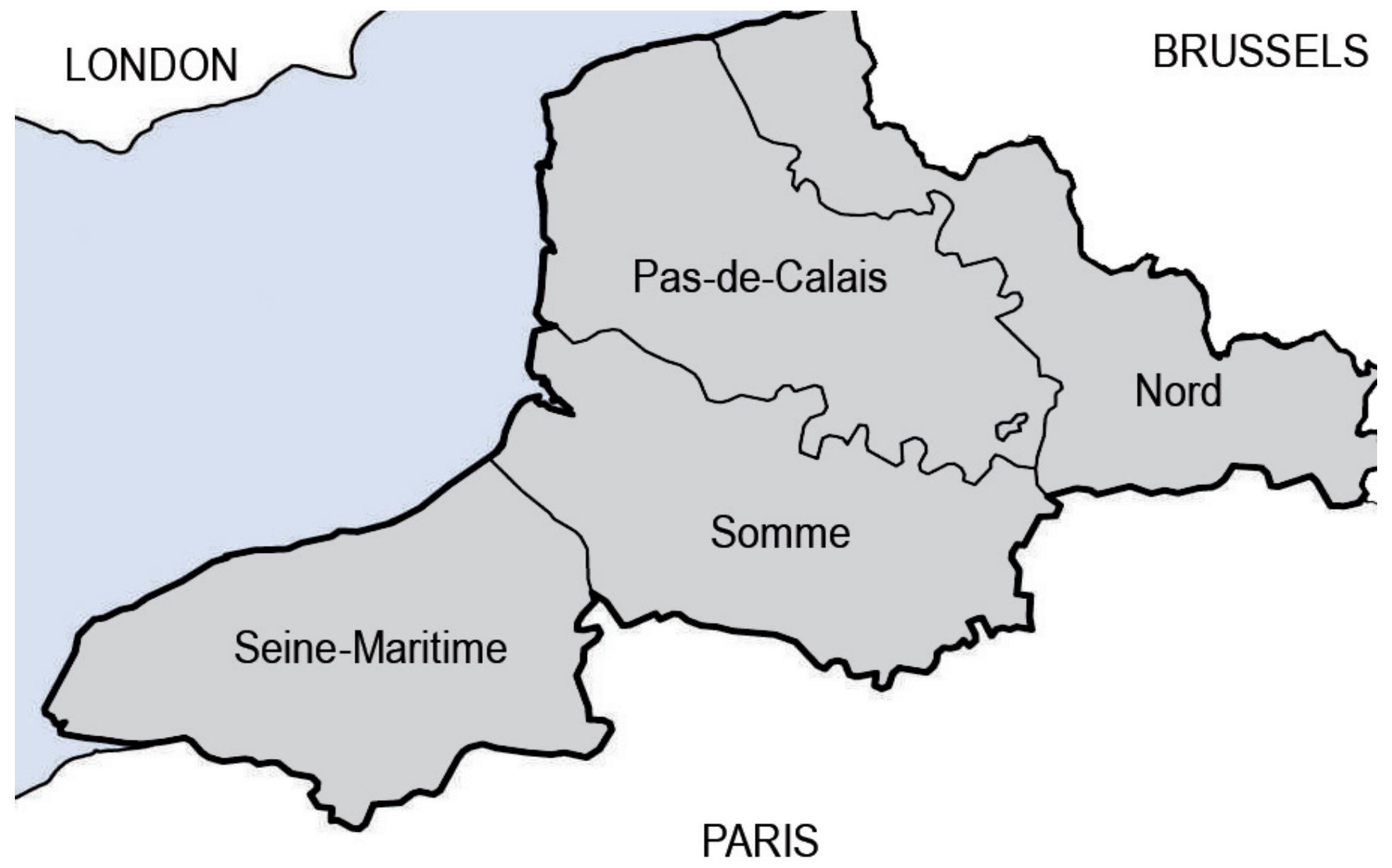
# Detecting Crohn's disease clusters using Spatial Scan Statistics

Alexandru Amărioarei[1,3], Michaël Genin[2], Corinne Gower[2], Cristian Preda[1,2], Manuela Sidoroff[3]

Université de Lille 1 (France), Université de Lille 2 (France), National Institute of R-D for Biological Sciences (Romania)

## Problem

Crohn's Disease(CD) is an inflammatory disease of the intestines, which has no known pharmaceutical or surgical cure. In addition, geographical variations of CD incidence have been reported worldwide reflecting putative variations in the distribution of environmental factors. In Northern France we were able to detect spatial heterogeneity in standardized incidence ratio (SIR) of CD (Declercq 2010). Between 1990 and 2006, 6 472 CD cases were recorded by the EPIMAD Registry of Northern France distributed in 273 cantons of Departments of Nord, Pas-de-Calais, Somme and Seine-Maritime (5 790 526 inhabitants).



## Spatial scan statistics model

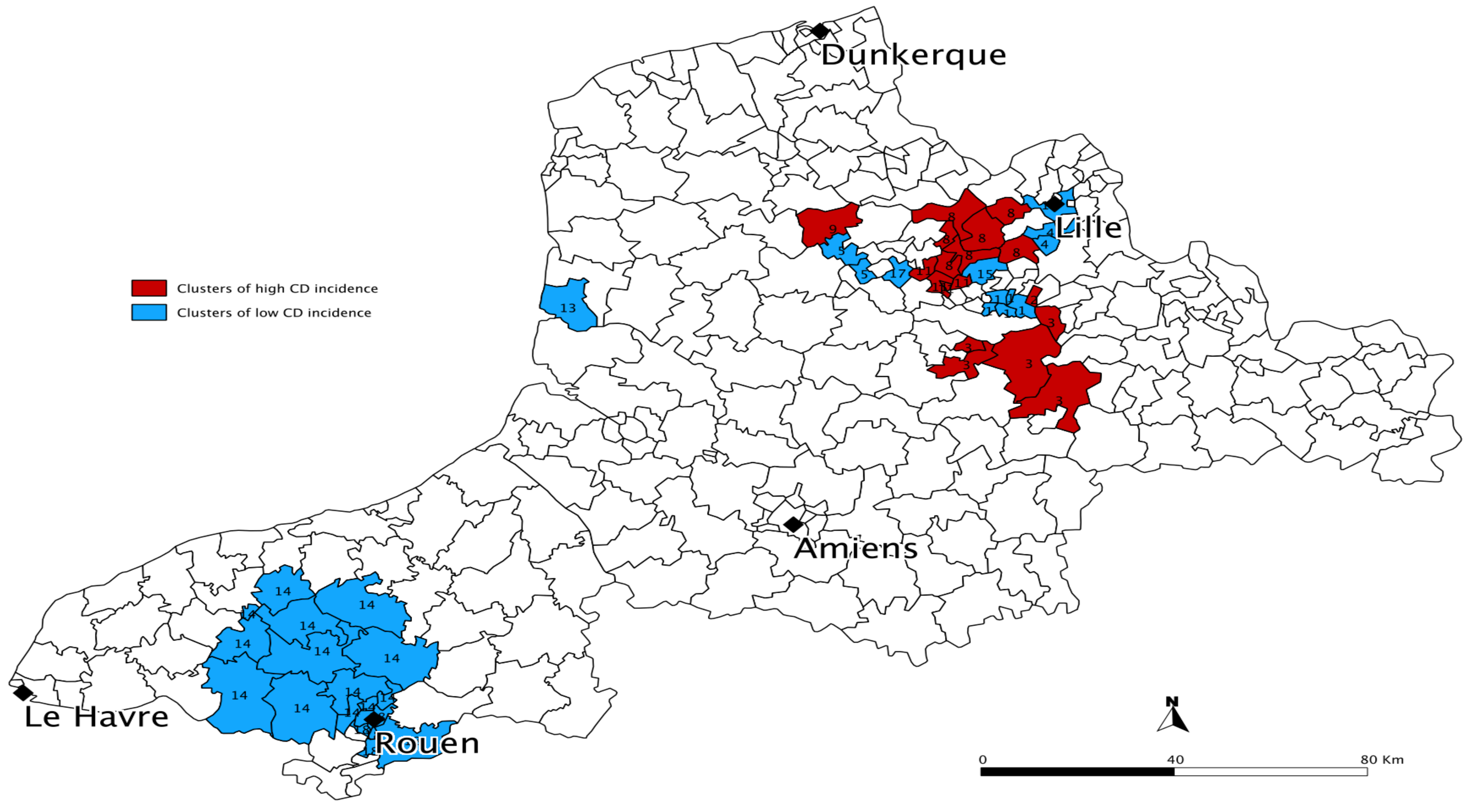
The spatial scan statistics method (Kulldorff 1995, 1997) was used to test for the presence of CD clusters and identify their approximate location. The following assumption is made: the number of CD cases in each canton is Poisson distributed. The method tests the null hypothesis that the risk of being affected by CD is constant throughout all cantons. It uses a 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), which moves across the area, using as center the centroid of the cantons. In total, we get a large number of circular windows which can candidate for being a cluster of CD cases, each containing a set of neighboring cantons.

Under the alternative hypothesis, there is at least one region for which the underlying risk is higher inside the region as compared to outside. 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). 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 $d_1, d_2, \dots, d_n$ 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 $L_0 = \frac{e^{-n_G}}{n_G!}\left(\frac{n_G}{\mu(G)}\right)^{n_G}\prod_{i=1}^{n}\mu(d_i)$ is the likelihood under the null hypothesis. The p-value, $\mathbb{P}(\nu > \nu_{obs})$, associated to the MLC is obtained based on Monte-Carlo random replications ($R = 9\,999$) under the null hypothesis. The calculations were performed using SaTScan®.


Dunkerque
Lille
Amiens
Le Havre
Rouen
🟥 Clusters of high CD incidence
🟦 Clusters of low CD incidence
13
9
8
5
17
15
4
3
1
14
N
0
40
80 Km

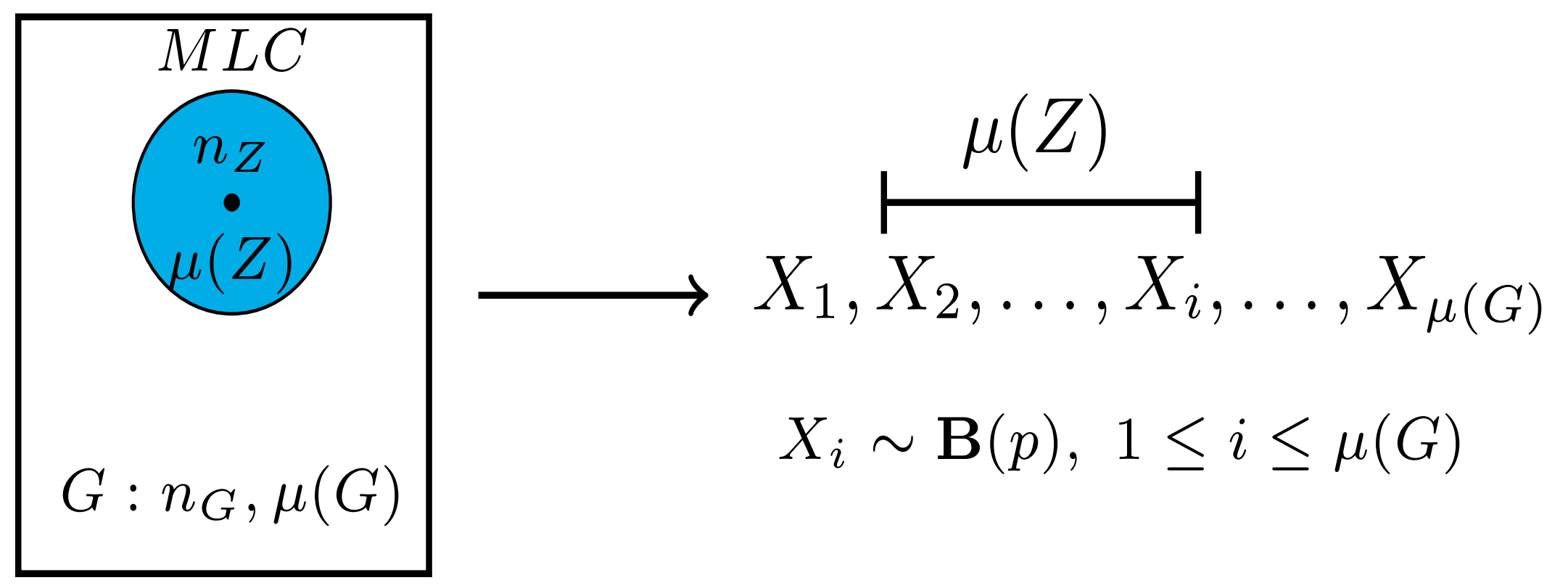
## Alternative approach for the Monte Carlo step

We observe that the spatial scan statistics method can be divided into two phases. The first phase, the detection, consists in finding the cluster that maximize the likelihood ratio (MLC) and the second phase, the inference, permits to test the significance of MLC based on Monte Carlo simulations. Depending on the problem, the last step, usually requires excessive computational time. To overcome this difficulty, an alternative to this phase is proposed.

We can model the population as a sequence $X_1, X_2, \dots, X_{\mu(G)}$ of i.i.d. Bernoulli trials, $\mathbb{P}(X_1 = 1) = p = 1 - \mathbb{P}(X_1 = 0)$, where 1 represents the presence of disease and 0 the absence. Assume that the sequence is scanned with a window of size $1 \leq \mu(Z) \leq \mu(G)$ and define the one dimensional scan statistic as

$$S_{\mu(Z)}(\mu(G)) = \max_{1\leq t\leq \mu(G)-\mu(Z)+1} \sum_{j=t}^{t+\mu(Z)-1} X_j.$$

The significance of the *most likely cluster* can then be tested by evaluating the tail probability $\mathbb{P}\left(S_{\mu(Z)}(\mu(G)) > n_Z\right)$.


$MLC$
$n_Z$
$\mu(Z)$
$G: n_G, \mu(G)$
$\mu(Z)$
$X_1, X_2, \dots, X_i, \dots, X_{\mu(G)}$
$X_i \sim \mathbf{B}(p),\ 1 \leq i \leq \mu(G)$

Several approximations have been proposed for the distribution of the one dimensional scan statistics. Taking $L = [\mu(G)/\mu(Z)]$ we have the following estimates:

(1) Naus 1982, using Markov like approximation, proposes the product type formula

$$\mathbb{P}\left(S_{\mu(Z)}(\mu(G)) \leq n_Z\right) \approx q_1\left(\frac{q_2}{q_1}\right)^{L-2}, \quad L > 2$$

(2) Haiman 2007, based on extreme value theory, shows that

$$\mathbb{P}\left(S_{\mu(Z)}(\mu(G)) \leq n_Z\right) \approx \frac{2q_1 - q_2}{(1 + q_1 - q_2 + 2(q_1 - q_2)^2)^L}$$

with a relative error of about $3.3L(1-q_1)^2$ and where

$$q_1 = \mathbb{P}\left(S_{\mu(Z)}(2\mu(Z)) \leq n_Z\right),\ q_2 = \mathbb{P}\left(S_{\mu(Z)}(3\mu(Z)) \leq n_Z\right).$$

The main advantage of the Bernoulli model consists in the fact that the probabilities $q_1$ and $q_2$ can be evaluated by exact formulas (Naus 1982). For the comparison of the two approaches, a simulation study was conducted. This study showed that the two methods provided consistent probabilities, the second approach presenting a superior power and a lower calculation time.

## Remarks

The study showed significant spatial heterogeneity of CD incidence in northern France during the period from 1990 to 2006, both confirming and extending previous data (Declercq et al. 2010). Using spatial ( and space-time) scan statistics, 14 spatial time constant clusters were identified. Among these clusters, 5 clusters of high incidence (total: 726 patients) and 9 clusters of low incidence (total:521) were detected. The existence of such clusters suggests that risk factors of CD are still at work in the studied region.

## References

[1] Declercq, C. et al.: Mapping of inflammatory bowel disease in northern France. *Inflamm Bowel Dis* **16** (2010), 807–812.

[2] Genin, M. et al.: Space-time clusters of Crohn's disease in northern France. *J Pub Med* (2013)