Méthodes Statistiques Appliquées à la l'analyse de Données Spatiales

Jean Gaudart

Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale

jean.gaudart@univmed.fr

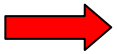
Faculté de Médecine
Université de la Méditerranée

plan

1. Introduction
   - 1.1 Données spatiales ?
   - 1.2 Nombreuses méthodes
2. Recherche d'agrégats
   - 2.1 Exemple: John Snow
   - 2.2 Classification des méthodes
   - 2.3 Hypothèse
3. Méthodes globales
   - 3.1 Moran
   - 3.2 Tango
   - 3.3 Poids et distances
4. Méthodes locales
   - 4.1 LISA
   - 4.2 Satscan

1. Introduction

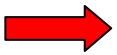
- **Données spatiales?**



– coordonnées géographiques

– relations entre variables : auto-corrélations spatiales

– fonction de la « distance »



– problèmes:

- relation spatiale d'un phénomène uniquement due à une variable spatialisée

  ex. VIH et conditions socio-économiques
- échelle spatiale
- hétérogénéité de population
- relation temporelle

  ex. grippe
- Données agrégées ⇒ études écologiques

- **Nombreuses méthodes selon :**

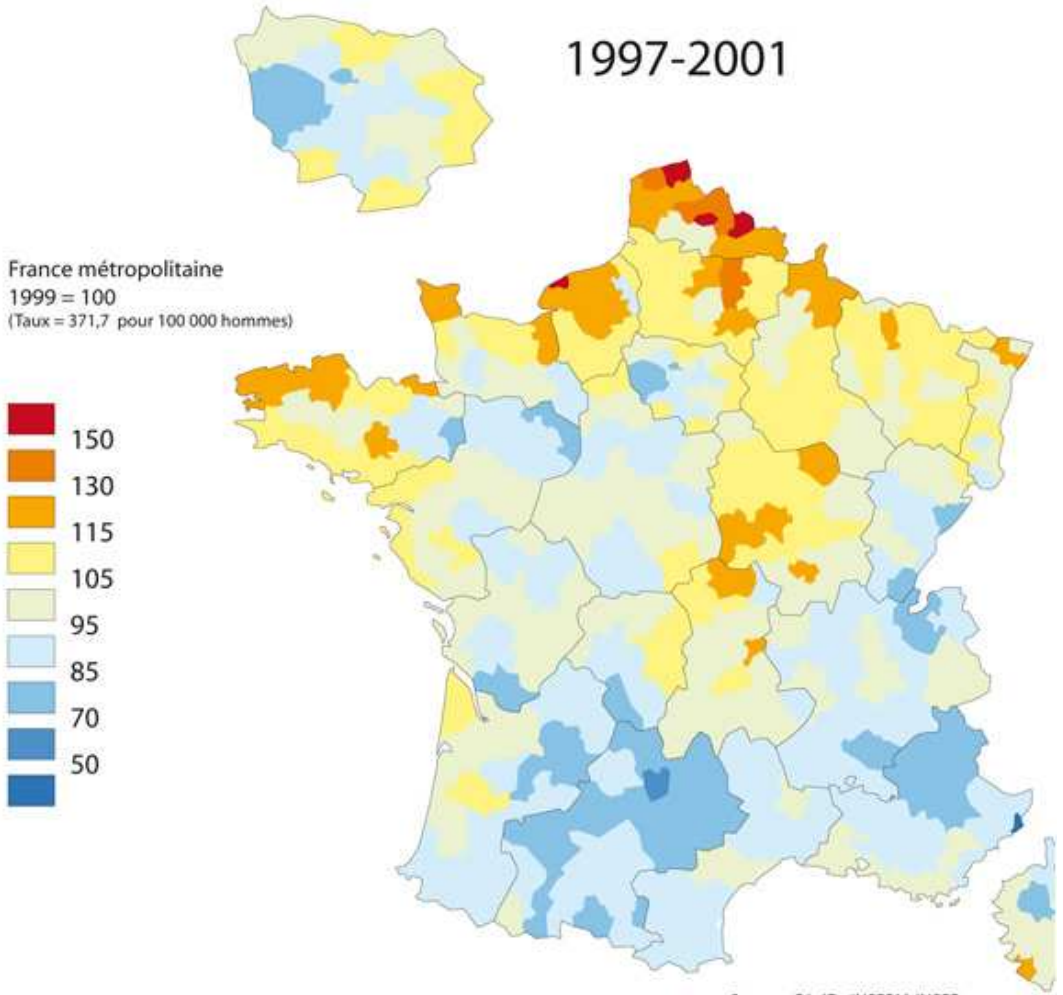
– Types de données, distributions particulières
  - variable continue (ex. NDVI)
  - nombre de cas (ex. nb de cas de VIH)

– Types de problématiques
  - construction de cartes
  - interpolation (« krigeage »)
  - recherche de tendances spatiales

    ex. axe cancer N – S en France
  - recherche d'agrégats

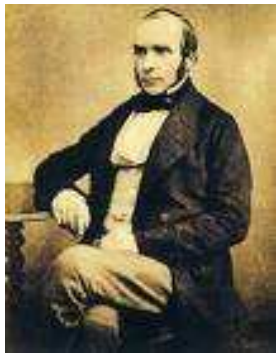
– Types d'hypothèses
  - distribution de la population
  - distribution des cas (uniforme, poisson, binomial etc)
  - cofacteurs

Ratios standardisés de mortalité par cancers
des hommes à l'échelle des zones d'emploi en France métropolitaine



2. Recherche d'agrégats - Clusters

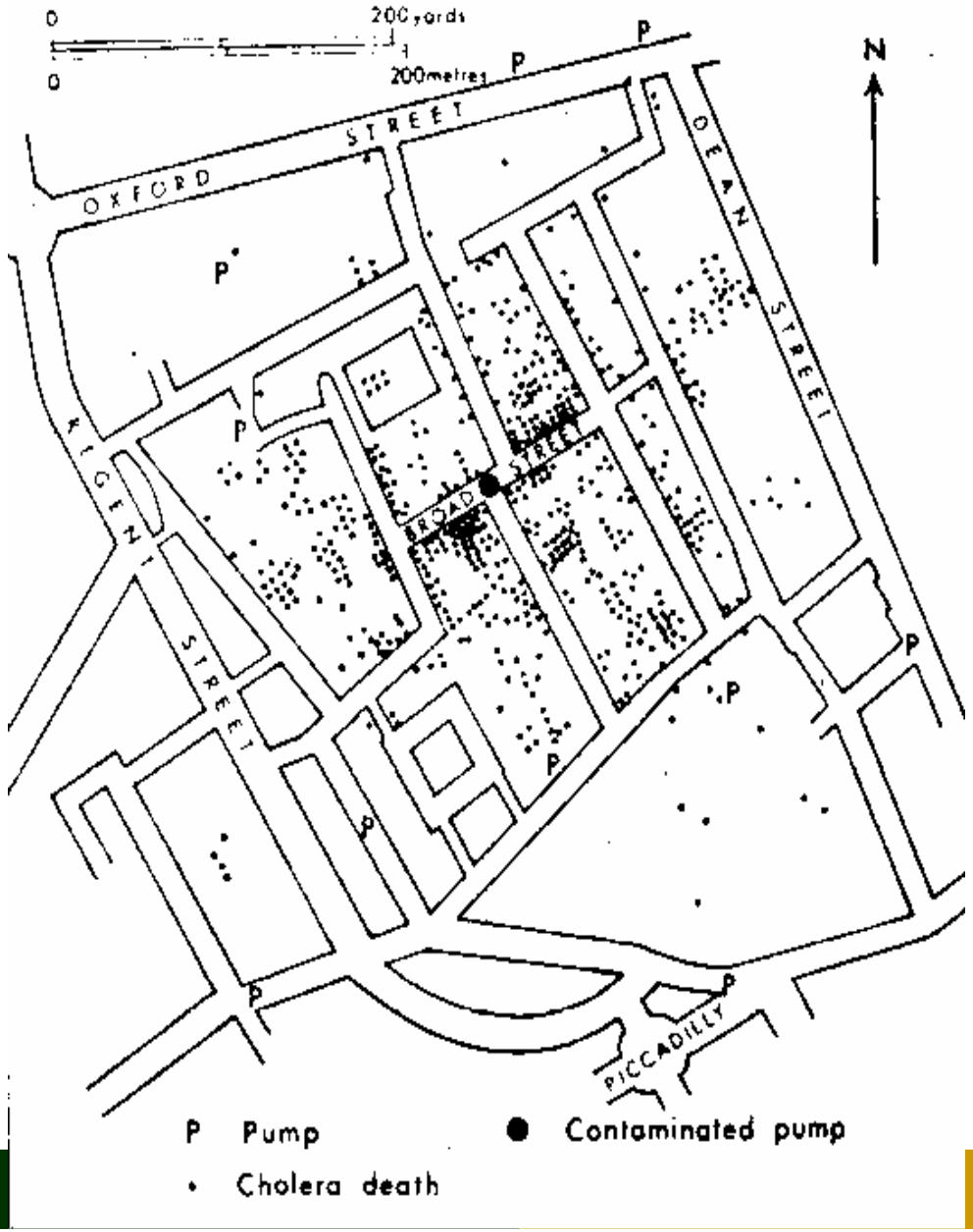
2.1 Exemple historique: John Snow



Anesthésiste anglais (1813-1858)

Épidémiologie du choléra (1854) :

Constatation : « à élévation égale, certains quartiers de Londres ont des taux d'incidence très différents »


0
200 yards
0
200 metres
N
OXFORD STREET
BROAD STREET
GREAT MARLBOROUGH STREET
REGENT STREET
PICCADILLY
P Pump
Cholera death
Contaminated pump

Source	Nombre d'habitations	Nombre de morts par choléra	Nombre de morts / 10000 habitations
en aval	40 046	1 263	315
en amont	26 107	98	37
reste de Londres	256 423	1 422	55

2.2 Classification

- autour d'un facteur de risque identifié, géo-localisé

  ex. centrales nucléaires / leucémies

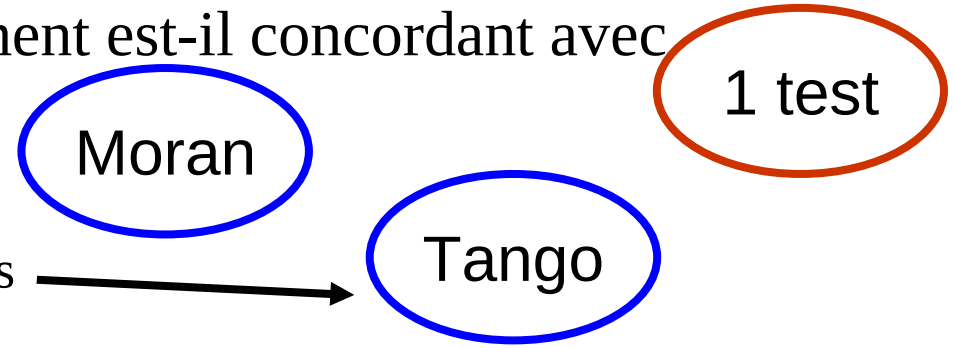
- Méthodes générales

  recherche d'agrégats sans facteur de risque géo-localisé

  - Globales – « clustering »
  - Locales

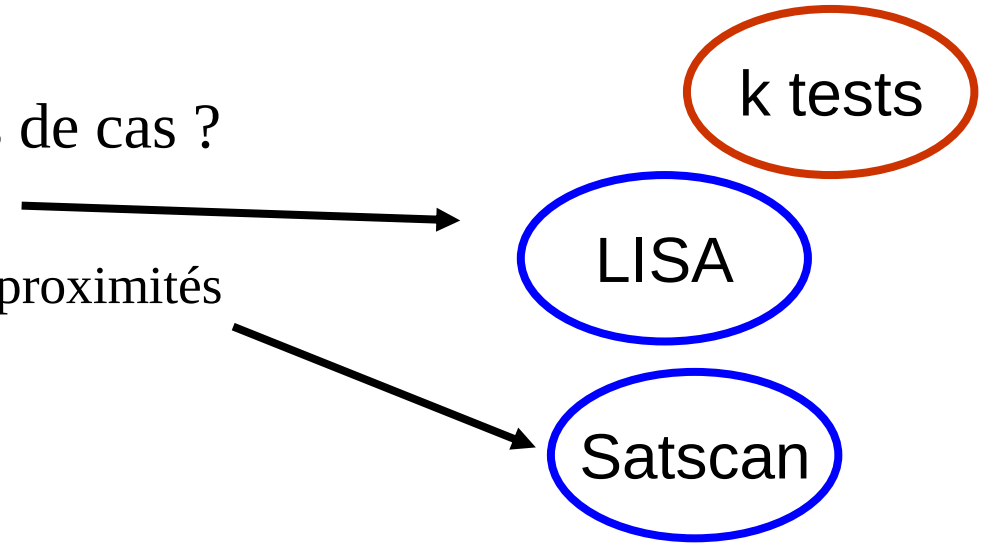
Détection globale

- Le pattern observé globalement est-il concordant avec l'hypothèse nulle ?
  - corrélations
  - comparaisons de distributions



Détection locale

- Y a-t-il localement un excès de cas ?
  - corrélations (voisinage)
  - comparaisons de groupes de proximités



2.3 Hypothèse

3 questions possibles :

- Les distributions des cas de chaque zone sont-elles indépendantes?
- Quelles sont ces distributions ?
- Le risque est-il constant sur l'ensemble de la zone géographique?

Hypothèses nulles possibles:

- Distribution des cas Uniforme sur le plan

Complete Spatial Randomness

- Distribution de Poisson hétérogène

Constant Risk Hypothesis

nombre de cas attendus dans la région *i*

$$E_i = \lambda n_i$$

Effectif région ***i***

risque constant

$$\tilde{\lambda} = \frac{O_+}{n_+}$$

nombre total de cas observés : $O_+ = \sum_{i=1}^{k} O_i$

effectif total : $n_+ = \sum_{i=1}^{k} n_i$

3. Méthodes Globales

3.1 Coefficient de Moran (1917-1988)

- Coefficient de corrélation
- Pondéré par les distances
- **Similarités** entre régions:

  écart à la moyenne de la région ***i***

  ⇔ écart à la moyenne de la région ***j***

$$I = \frac{K \times \sum_{i,j}^{n} w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sqrt{\sum_{i=1}^{n} w_{+} \times \sum_{i=1}^{K} (Y_i - \bar{Y})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

Plus les zones i et j sont éloignées, moins le poids est important

- *Application*

Fichier de données: DATA1.xls

num : numéros de l'unité spatiale

cas : nombre de malades

n : taille de la population à risque

x, *y* : coordonnées

date : date de l'étude

A transformer en .csv !!

- *Application: préparation pour l'analyse*

```
rm(list=ls(all=TRUE))
 local({pkg <- select.list(sort(.packages(all.available = TRUE)))
+ if(nchar(pkg)) library(pkg, character.only=TRUE)})
```

remove

charger le package

```
SPAT<-read.csv2("D:\\OPT17\\DATA1.csv",header=TRUE)
attach(spat)

TAB<-data.frame(Observed=cas)
TAB<-cbind(TAB, Expected=n*sum(cas)/sum(n))
```

charger les données

construire le tableau de données adapté

```
SPAT.xy = SPAT[c("x", "y")]
coordinates(SPAT.xy) <- ~x+y
class(SPAT.xy)

coords<-coordinates(SPAT.xy)
```

construire le tableau des coordonnées

construire le tableau des distances et la matrice des poids

```
dlist<-dnearneigh(coords, 0, Inf)
dlist<-include.self(dlist)
dlist.d<-nbdists(dlist, coords)
col.W<-nb2listw(dlist,glist=lapply(dlist.d,function(x){exp(-x)}),style="C")
```

- *Application: estimation du coefficient de Moran*

```
I<-moranI.stat(TAB, listw=col.W, n=length(dlist), S0=Szero(col.W) )
```

données
{Observées, Attendues}

pondérations
fonction des distances
entre les unités spatiales

effectif

somme totale des poids

$$S0 = w_+ = \sum_{i=1}^{k} w_i$$

```
I
[1] 0.3441209
```

- Test:

H0: ***I=0***; malades spatialement indépendants

H1: ***I>0***

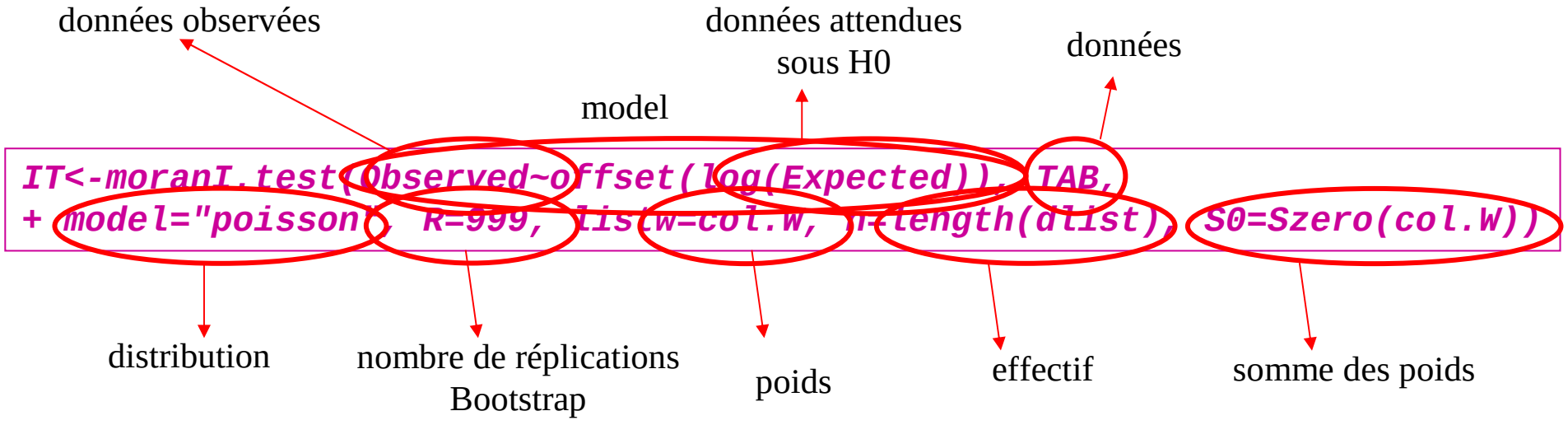
– Sous H0: $z = \frac{I - E(I)}{\sqrt{\mathrm{var}(I)}} \quad \sim> \mathcal{N}(0;1)$

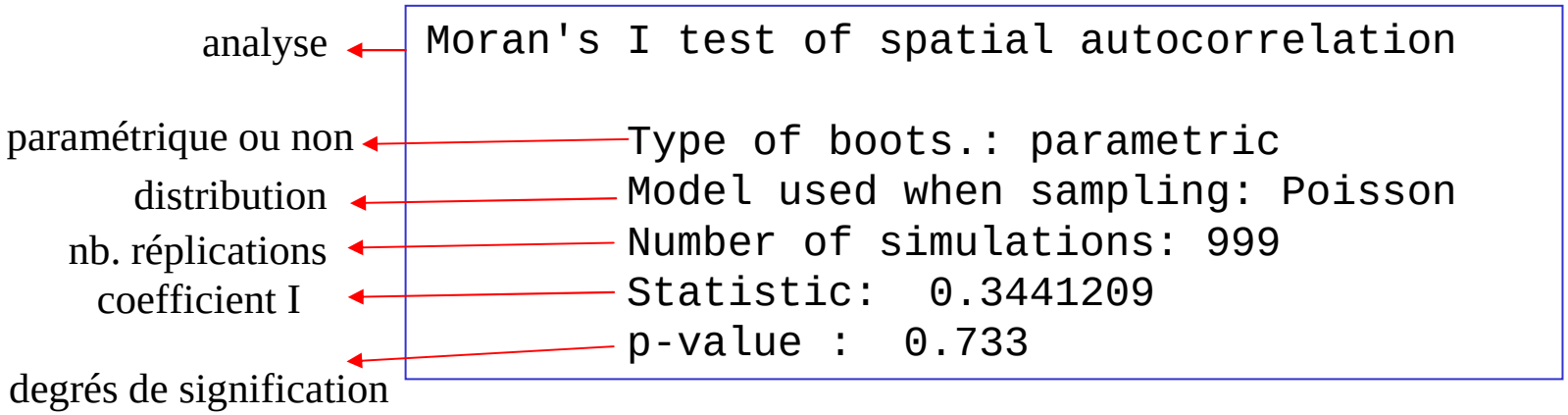
– Condition : distribution des ***Y*** normale

- insoutenable
- unité statistique: région, avec k petit

– Loi de ***I*** inconnue ⇒ Monte-Carlo ou Bootstrap

- *Application: test du coefficient de Moran*

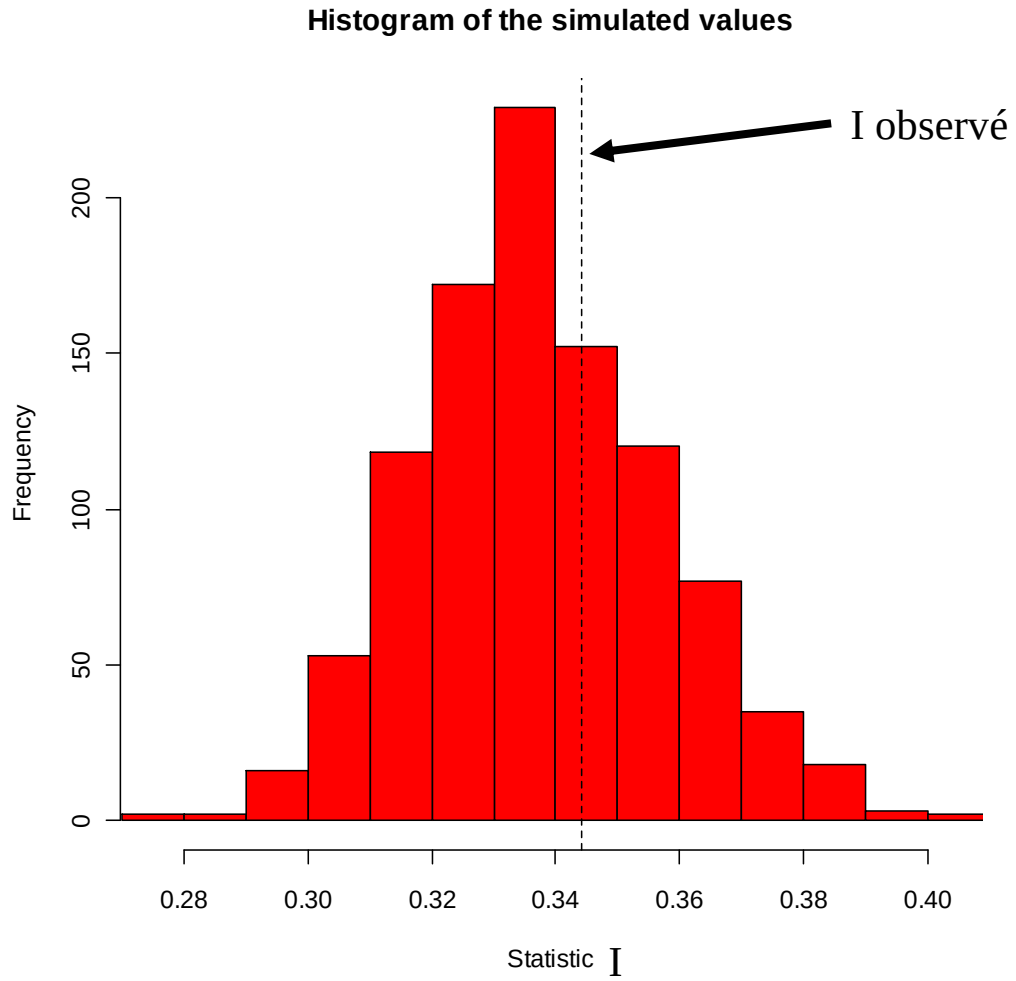




par défaut: *alternative="greater "*

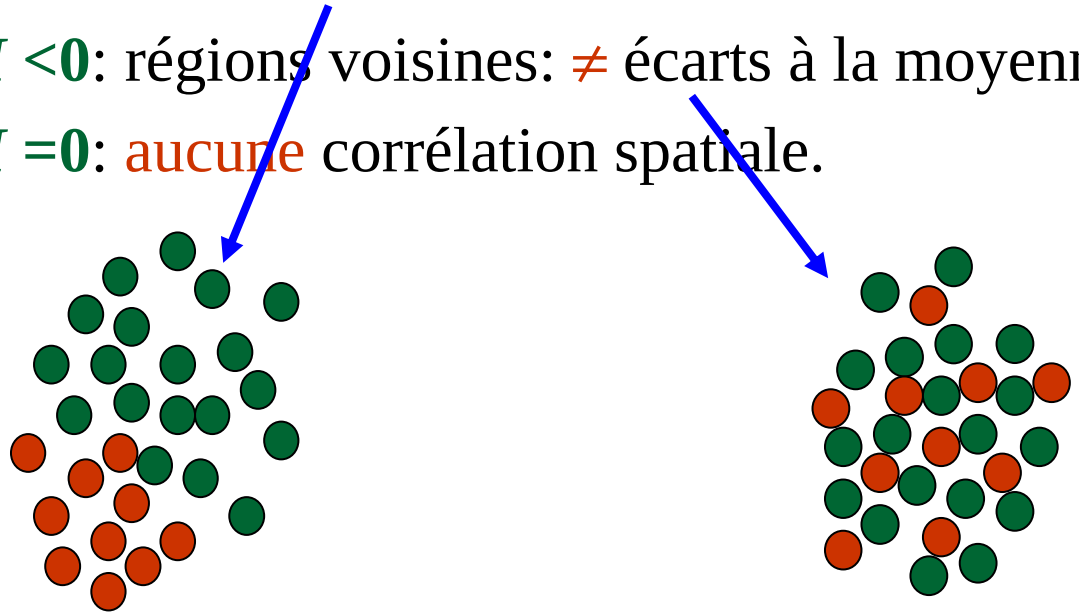
- *Application: test du coefficient de Moran (suite)*

```
plot(IT)
```



- Interprétation :

- ***I* >0**: régions voisines: mêmes écarts à la moyenne = pattern sous forme de clusters
- ***I* <0**: régions voisines: ≠ écarts à la moyenne, = pattern régulier.
- ***I* =0**: aucune corrélation spatiale.



- Mesure de l'écart à la moyenne générale:

pas d'interprétation locale possible.

3.2 Statistique de Tango

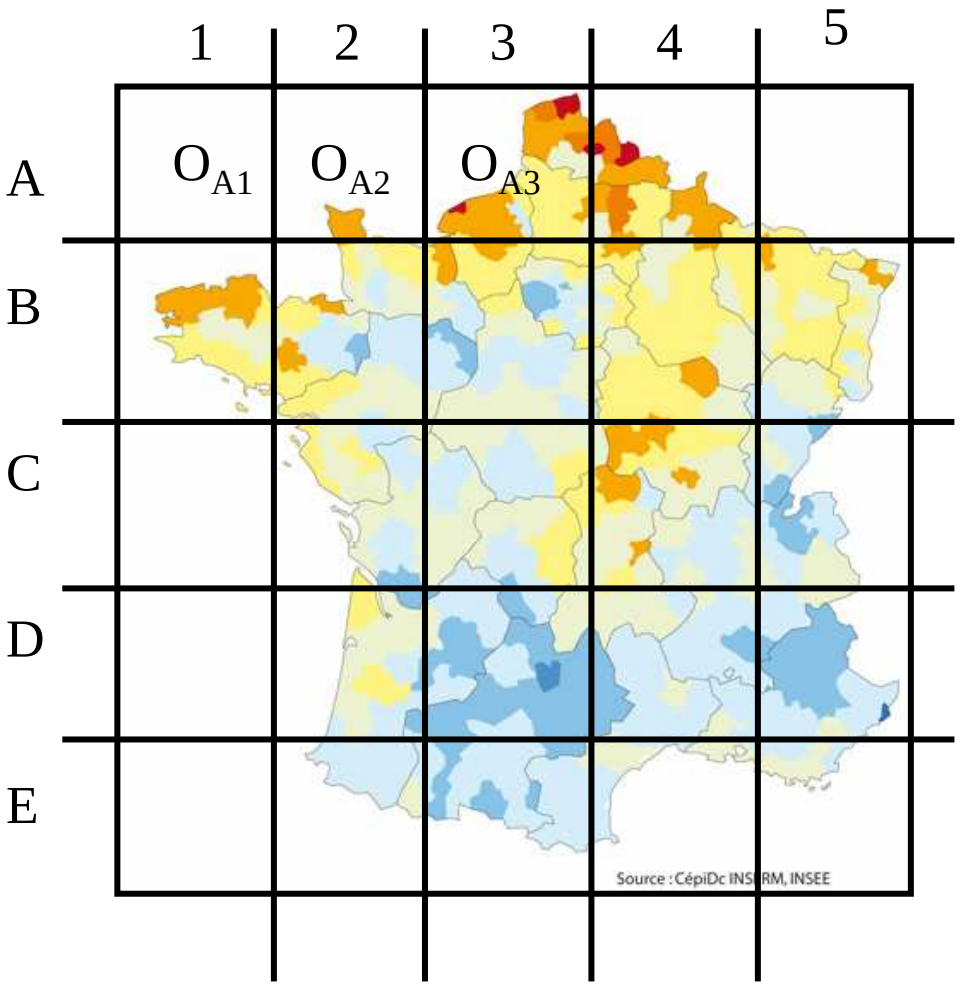
- Principe

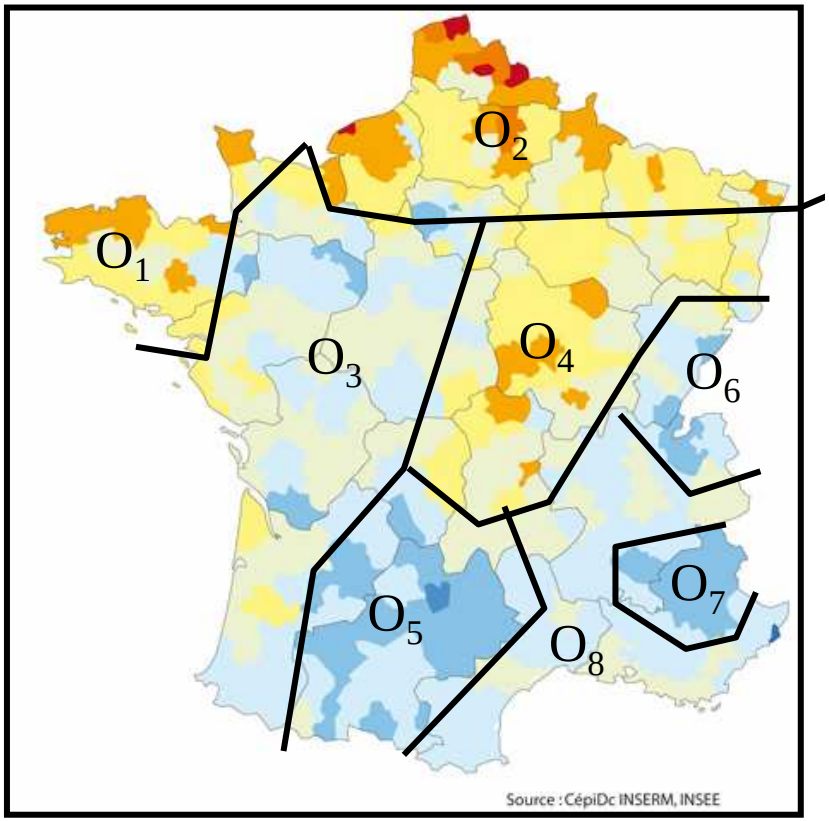
Observés

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - C_i)^2}{C_i}$$

Attendus

	1	2	
A	O_{A1}	O_{A2}	O_{A+}
B	O_{B1}	O_{B2}	O_{B+}
	O_{1+}	O_{2+}	O_{+}


1
2
3
4
5
A
B
C
D
E
O_{A1}
O_{A2}
O_{A3}
Source : CépiDc INSERM, INSEE


O_1
O_2
O_3
O_4
O_5
O_6
O_7
O_8
Source : CépiDc INSERM, INSEE

3.2 Statistique de Tango

- Principe

Observés

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - C_i)^2}{C_i}$$

Attendus

$$C_i = \tilde{\lambda} n_i$$

	1	2	
A	O_{A1}	O_{A2}	O_{A+}
B	O_{B1}	O_{B2}	O_{B+}
	O_{1+}	O_{2+}	O_{+}

- Généralisation spatiale du χ^2

Poids / distances

$$T = \sum_{i,j}^{K} w_{ij} \left(\frac{O_i}{O_+} - \frac{n_i}{n_+} \right) \left(\frac{O_j}{O_+} - \frac{n_j}{n_+} \right)$$

effectif zone j

effectif zone i

observation zone i

observation zone j

Plus les zones i et j sont éloignées, moins le poids est important

- *Application: estimation de la statistique de Tango*

```
T<-tango.stat(TAB, col.W, zero.policy=TRUE)
```

données
{Observées, Attendues}

pondérations
fonction des distances
entre les unités spatiales

```
  T
[1] 0.0004400708
```

- Test:

H0: pas d'écart entre observé et théorique

H1: il y a des zones différentes

- Sous H0, T~>$\mathcal{N}$(E(t);var(T))

Mais k petit et convergence lente

- Approximation: sous H0, ou MC ou Bootstrap

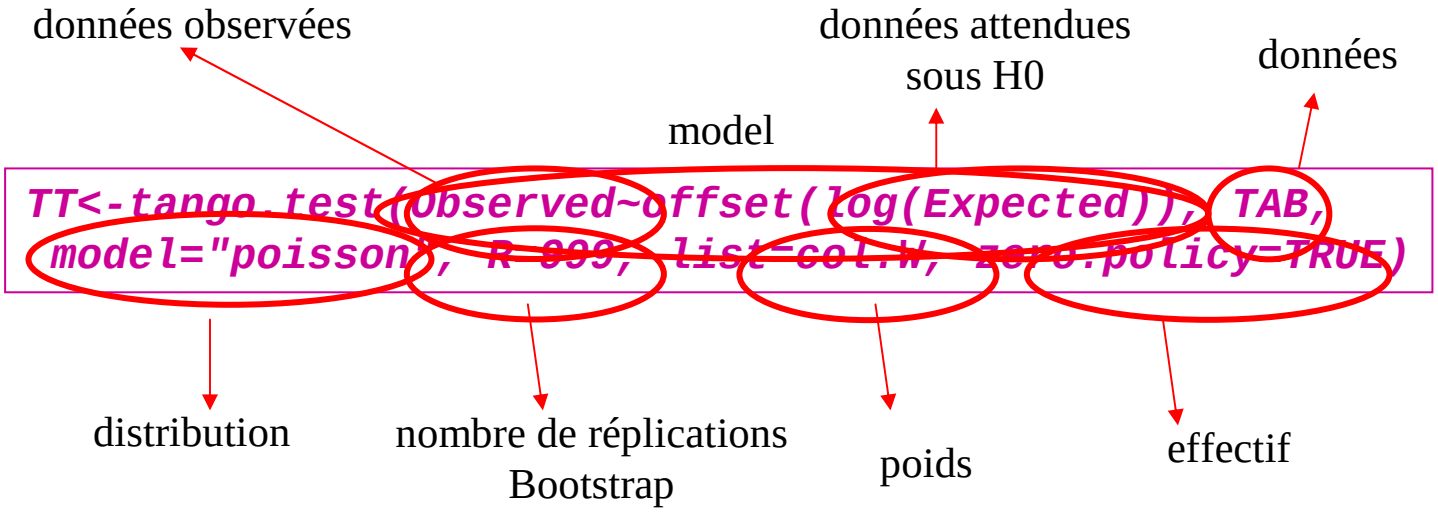
$$\nu + \frac{T - E(T)}{\sqrt{\mathrm{var}(T)}}\sqrt{2\nu} \xrightarrow{a} \chi^2_\nu$$

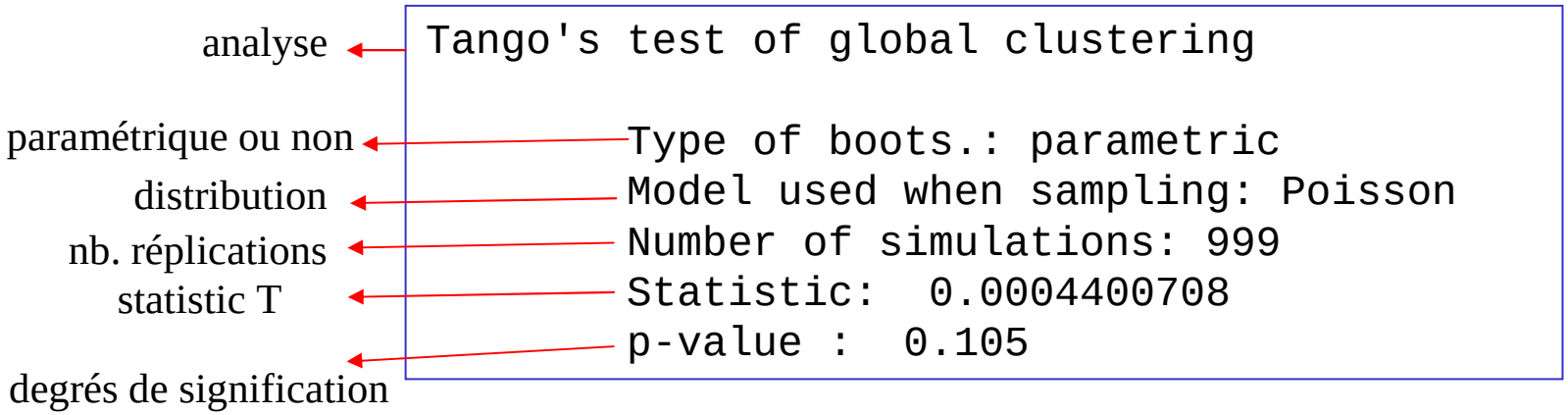
- Test de comparaison de distributions

+ autocorrélations spatiales

$$T = \sum_i^K w_{ii}\left(\frac{O_i}{O_+} - \frac{n_i}{n_+}\right)^2 + \sum_{i,j\neq i}^K w_{ij}\left(\frac{O_i}{O_+} - \frac{n_i}{n_+}\right)\left(\frac{O_j}{O_+} - \frac{n_j}{n_+}\right)$$

- *Application: test de la statistique de Tango*

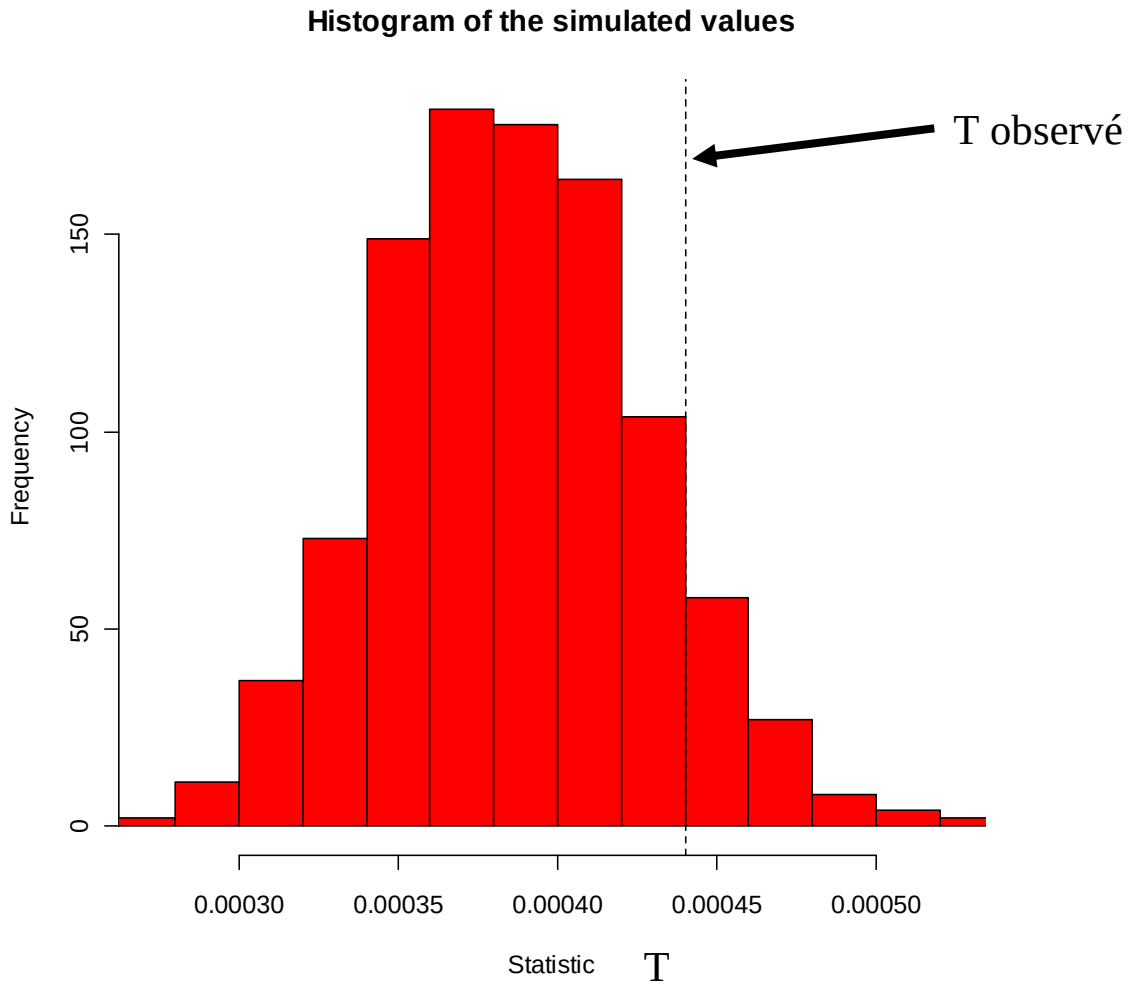




par défaut: *alternative="greater "*

- *Application: test de la statistique de Tango (suite)*

```
plot(TT)
```



3.3 Distances et Poids

- Distances entre localisations (ou centres des régions)

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2}$$

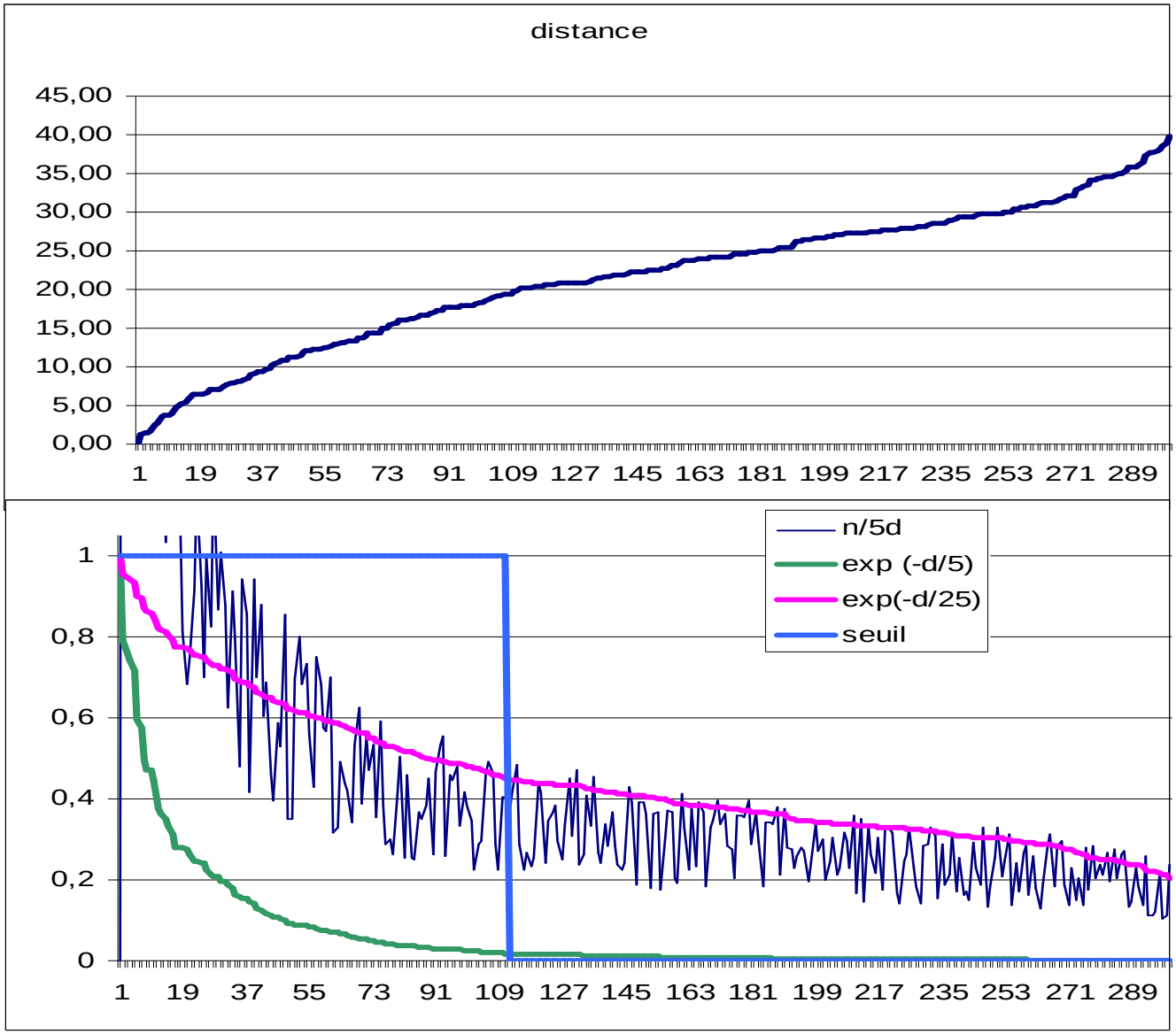
- Différents poids = différents résultats !!

$$w_{ij} = \begin{cases} 1, si\ d_{ij} < \delta \\ 0, si\ non \end{cases}$$

$$w_{ij} = e^{\left(-\frac{d_{ij}}{\tau}\right)}$$

$$w_{ij} = d_{ij} / n_v$$

Exemple:
Pour le point le plus à l'ouest



4. Méthodes Locales

4.1 LISA de Anselin

Local Indicator of Spatial Autocorrelation

Application locale du coefficient de Moran

A proximité d'un cas observé, les cas sont-ils regroupés?

Global

$$I = \frac{K \times \sum_{i,j} w_{ij} \left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)}{w_+ \times \sum_{i=1}^{K} \left(Y_i - \bar{Y}\right)^2}$$

Local: pour la région i

$$I_i = \left(Y_i - \bar{Y}\right) \times \sum_{j=1}^{K} w_{ij} \left(Y_j - \bar{Y}\right)$$

poids

Ecart de la région i à la moyenne

Ecart des autres régions à la moyenne

- Test: Pour chaque région

H0: $\boldsymbol{I_i=0}$ indépendance des régions/ voisins

H1: $\boldsymbol{I_i>0}$

– Sous H0 et condition $\mathcal{N}$: $z = \frac{I_i - E(I_i)}{\sqrt{\mathrm{var}(I_i)}}$ ~>$\mathcal{N}(0;1)$

➡ Condition insoutenable:

=>Loi de $\boldsymbol{I_i}$ inconnue => Monte-Carlo ou Bootstrap

➡ – **Tests multiples et corrélés**

=> *correction de Bonferroni* $\alpha_i = \frac{\alpha}{n_v}$

- *Application: test de LISA*

```
dlist2<-dnearneigh(coords, 0, Inf)
dlist2.d<-nbdists(dlist2, coords)
```

construire le tableau des distances et la matrice des poids

données observées

```
LM<-localmoran(TAB$Observed, nb2listw(dlist2),
     p.adjust.method="bonferroni")
```

ajustement

poids

estimation

p-value

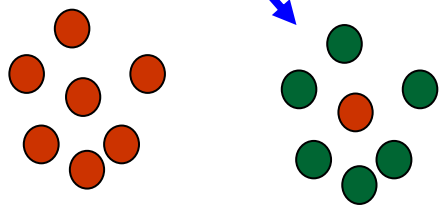
```
          Ii          E.Ii  Var.Ii     Z.Ii    Pr(z > 0)
  [1,] -4.329e-03 -0.003 1.721e-05 -0.237   1
  [2,] -5.324e-03 -0.003 1.721e-05 -0.47    1
  [3,] -1.009e-02 -0.003 1.721e-05 -1.63    1

...
```

- Interprétation

– $I_i > 0$: régions voisines similaires à région ***i***
= cluster local

– $I_i < 0$: régions voisines: ≠ à région ***i*** ,
= région ***i*** particulière.

– $I_i = 0$: aucune corrélation spatiale entre la région ***i*** et ses voisins.



$$\sum_{i=1}^{k} I_i \propto I_{global}$$

4.2 Satscan de Kulldorf

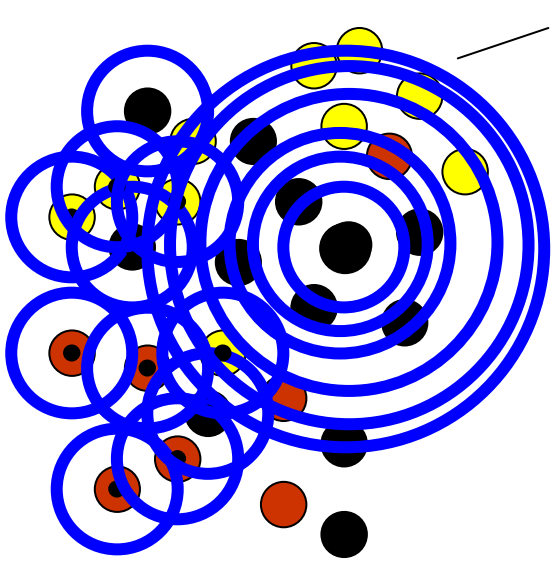
- Principe:
  - Fenêtre circulaire,
  - Rayon variable, centre variable (point ou centre)
  - Balaye le plan

  => clusters potentiels => statistique

$$T_k \propto \max_{nf} \left(\frac{O_{\text{int}}}{C_{\text{int}}} \right)^{O_{\text{int}}} \left(\frac{O_{ext}}{C_{ext}} \right)^{O_{ext}}$$

avec

$$C_i = \tilde{\lambda} n_i$$

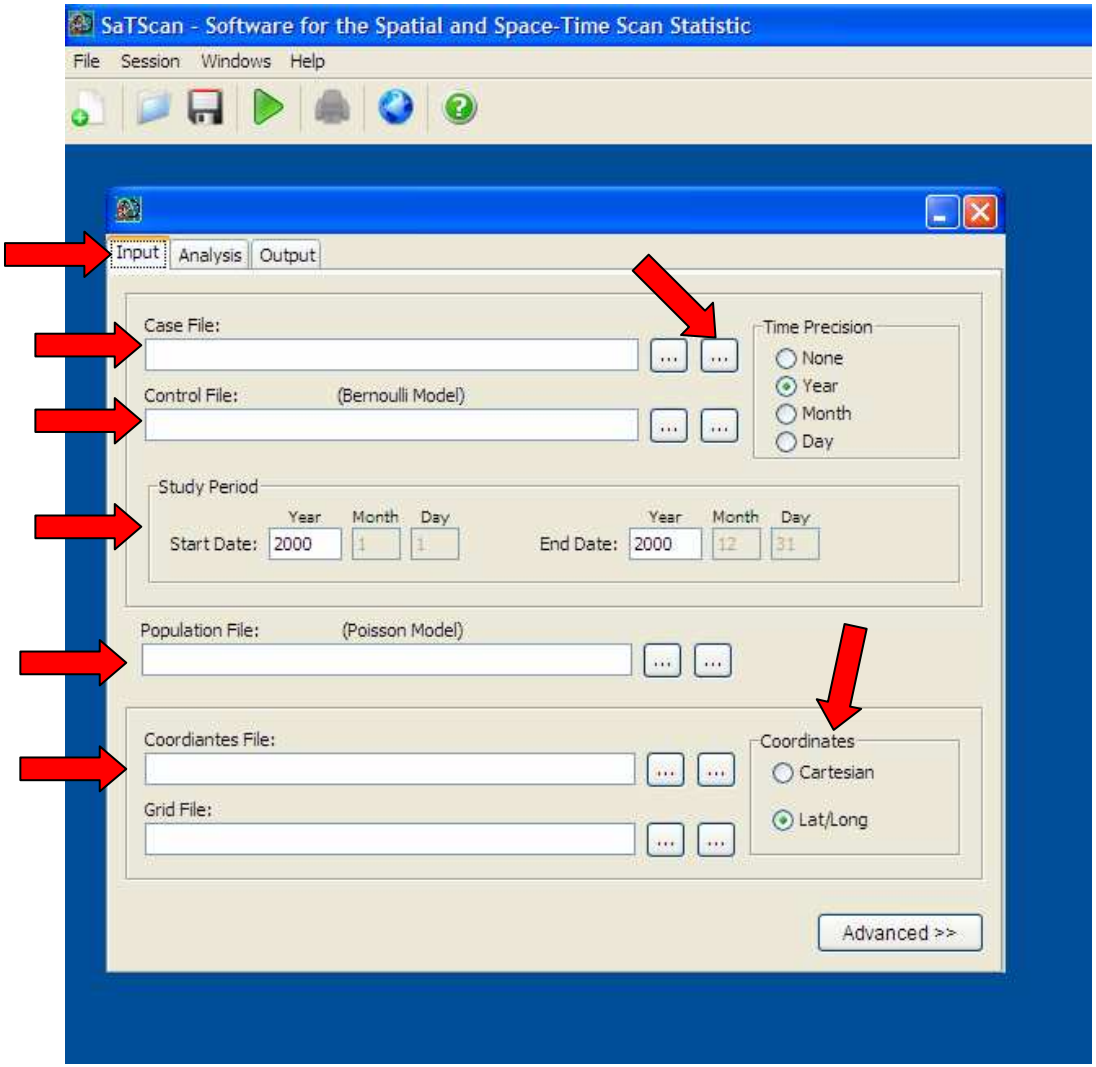


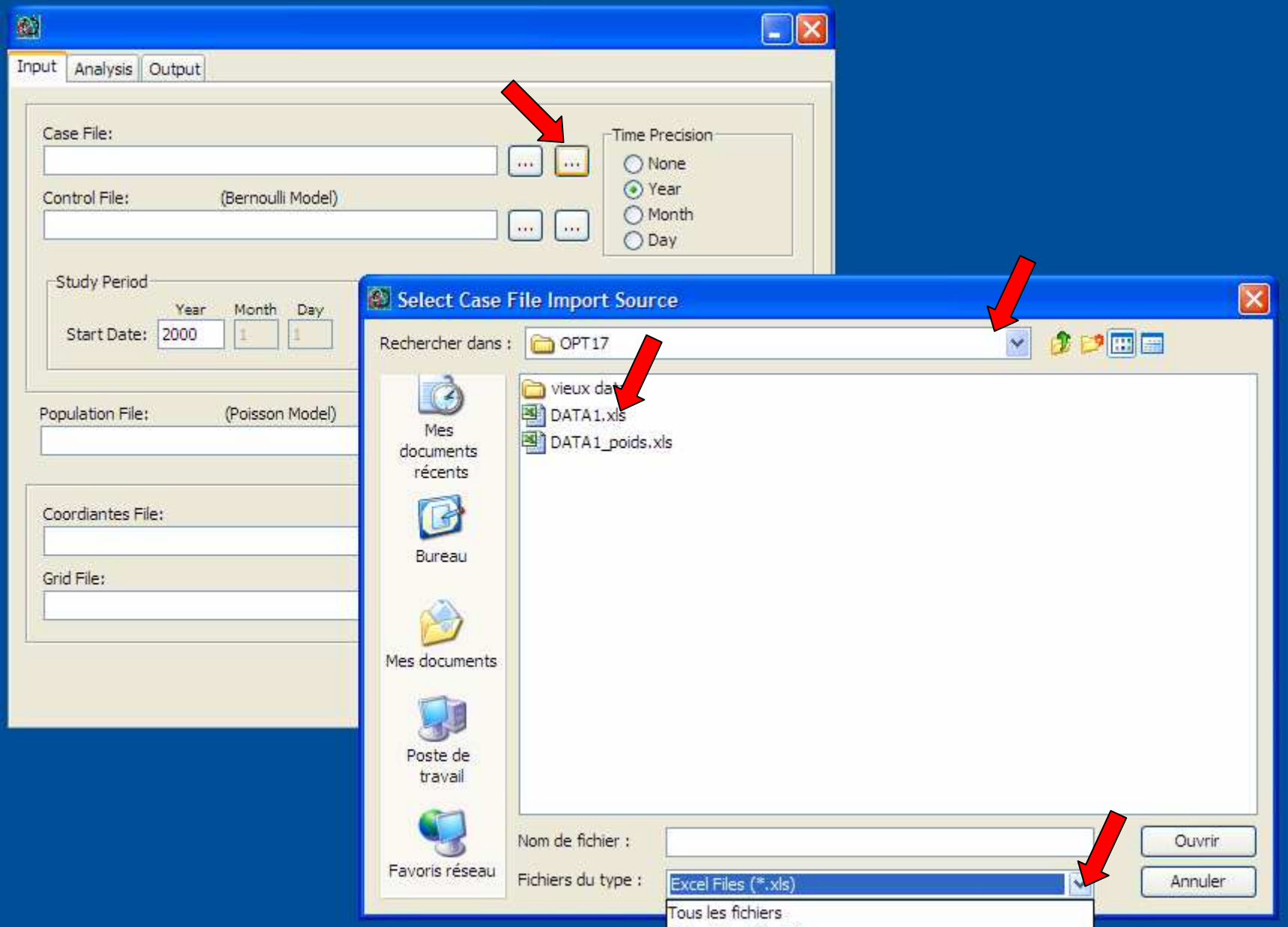
$$\left(\frac{O_{int}}{C_{int}}\right)^{O_{int}} \left(\frac{O_{ext}}{C_{ext}}\right)^{O_{ext}}$$

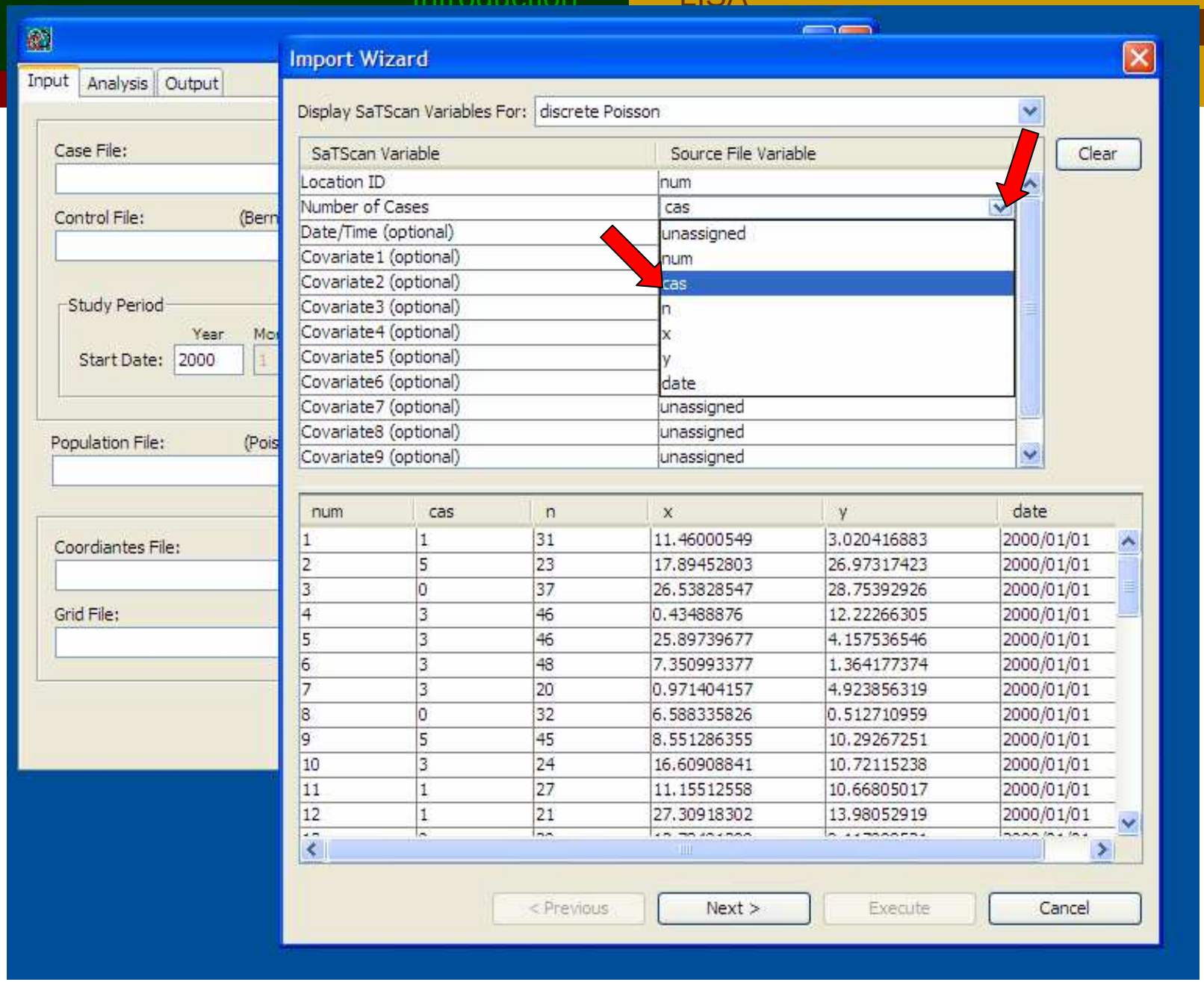
- Loi de T_k inconnue => Monte-Carlo

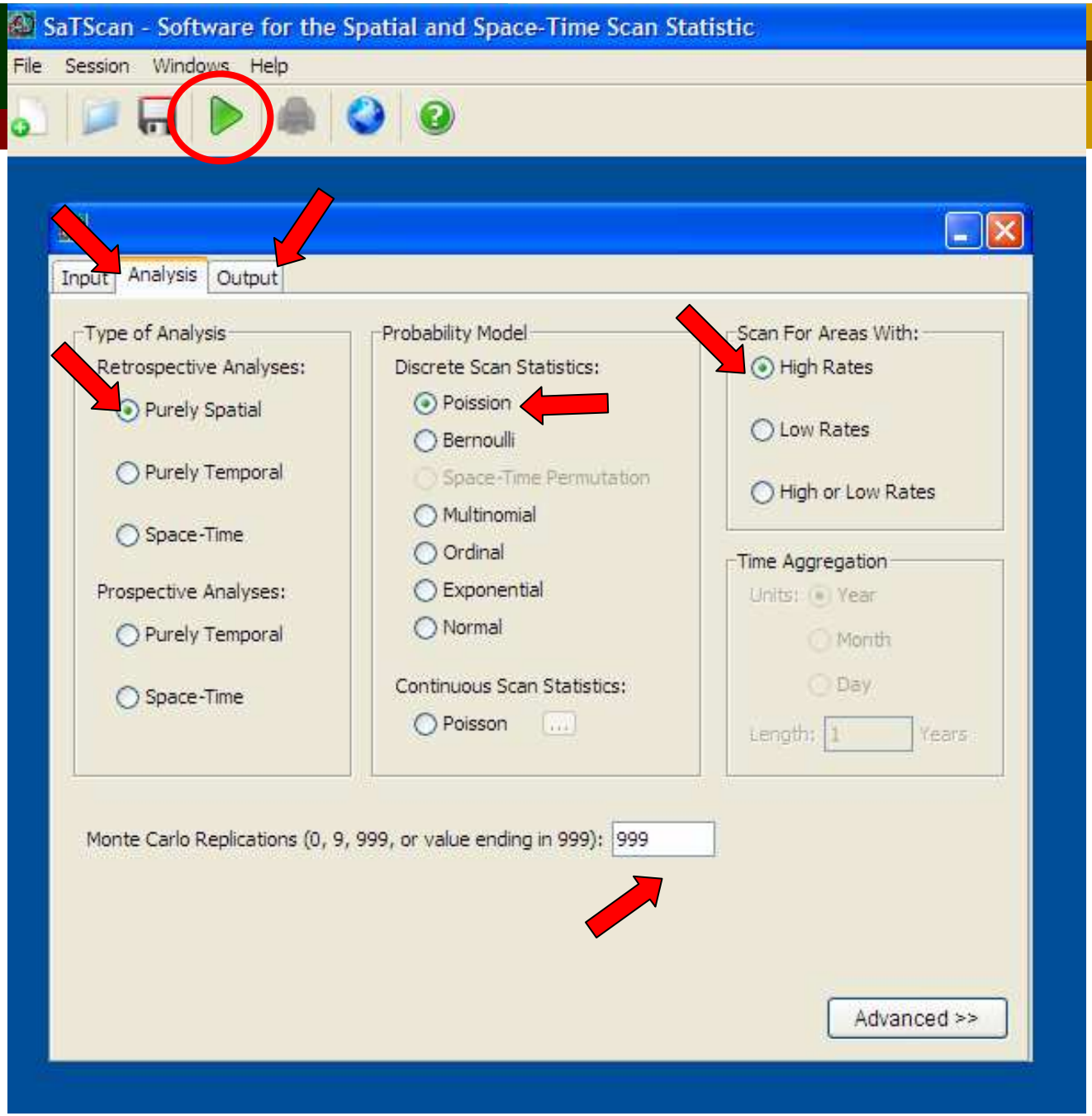
- Interprétation:
  - cluster: excès de cas dans la fenêtre
  - $\frac{O_i}{C_i}$ risque relatif ou rapport d'incidence

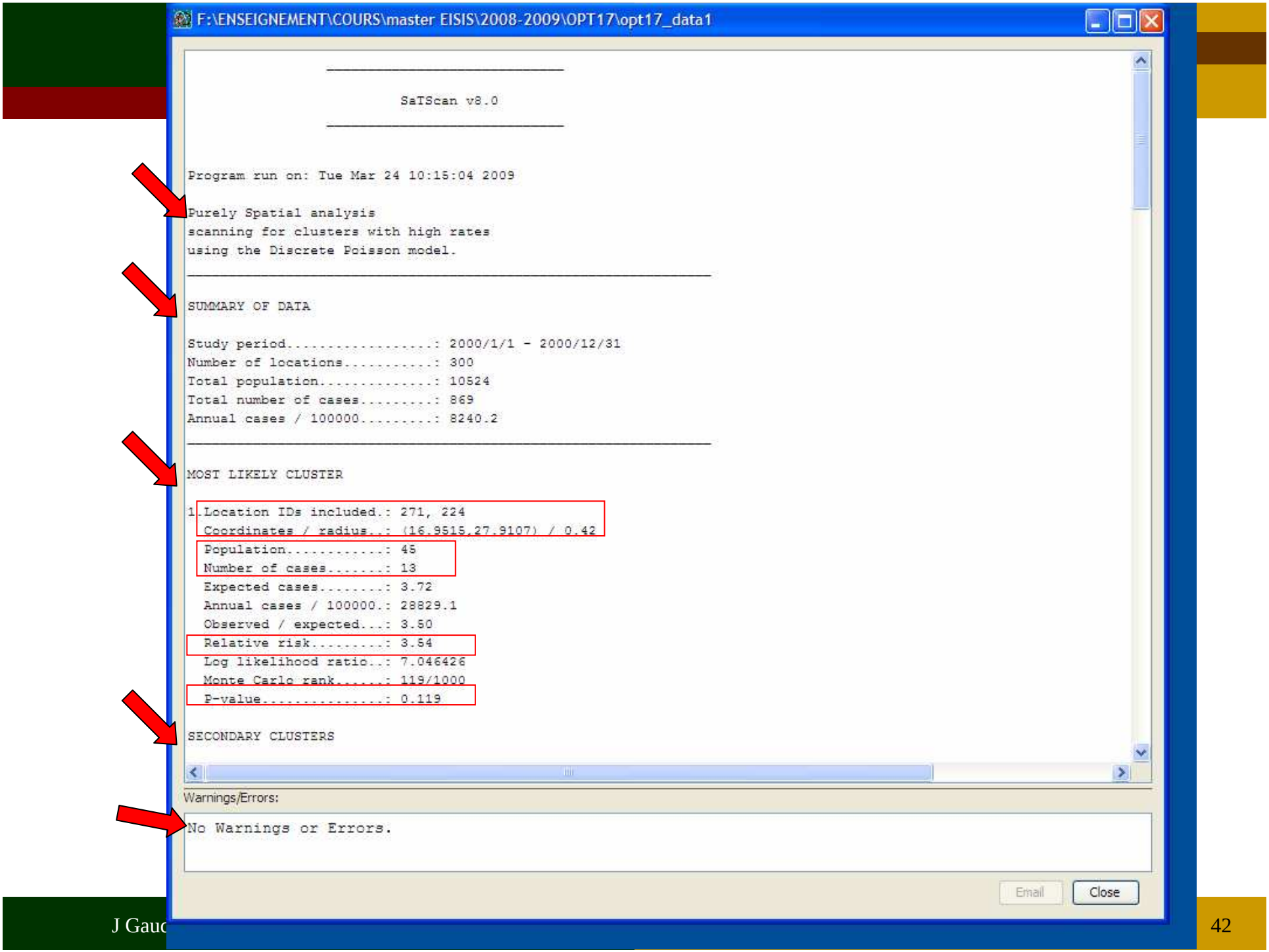
- *Application: logiciel SATSCAN* http://www.satscan.org




Input
Analysis
Output
Case File:
Control File:
(Bernoulli Model)
Time Precision
None
Year
Month
Day
Study Period
Year
Month
Day
Start Date:
2000
Population File:
(Poisson Model)
Coordiantes File:
Grid File:
Select Case File Import Source
Rechercher dans :
OPT17
vieux da
DATA1.xls
DATA1_poids.xls
Mes documents récents
Bureau
Mes documents
Poste de travail
Favoris réseau
Nom de fichier :
Fichiers du type :
Excel Files (*.xls)
Ouvrir
Annuler
Tous les fichiers
dBase Files (*.dbf)
Delimited Files (*.csv)
Excel Files (*.xls)
Text Files (*.txt)
Case Files (*.cas)


Input
Analysis
Output
Case File:
Control File:
Study Period
Year
Start Date:
2000
Population File:
Coordiantes File:
Grid File:
Import Wizard
Display SaTScan Variables For:
discrete Poisson
Clear
SaTScan Variable
Source File Variable
Location ID
num
Number of Cases
cas
Date/Time (optional)
Covariate1 (optional)
Covariate2 (optional)
Covariate3 (optional)
Covariate4 (optional)
Covariate5 (optional)
Covariate6 (optional)
Covariate7 (optional)
Covariate8 (optional)
Covariate9 (optional)
unassigned
num
cas
n
x
y
date
unassigned
unassigned
unassigned
num cas n x y date
1 1 31 11.46000549 3.020416883 2000/01/01
2 5 23 17.89452803 26.97317423 2000/01/01
3 0 37 26.53828547 28.75392926 2000/01/01
4 3 46 0.43488876 12.22266305 2000/01/01
5 3 46 25.89739677 4.157536546 2000/01/01
6 3 48 7.350993377 1.364177374 2000/01/01
7 3 20 0.971404157 4.923856319 2000/01/01
8 0 32 6.588335826 0.512710959 2000/01/01
9 5 45 8.551286355 10.29267251 2000/01/01
10 3 24 16.60908841 10.72115238 2000/01/01
11 1 27 11.15512558 10.66805017 2000/01/01
12 1 21 27.30918302 13.98052919 2000/01/01
< Previous
Next >
Execute
Cancel


SaTScan - Software for the Spatial and Space-Time Scan Statistic
File Session Windows Help
Input Analysis Output
Type of Analysis
Retrospective Analyses:
Purely Spatial
Purely Temporal
Space-Time
Prospective Analyses:
Purely Temporal
Space-Time
Probability Model
Discrete Scan Statistics:
Poission
Bernoulli
Space-Time Permutation
Multinomial
Ordinal
Exponential
Normal
Continuous Scan Statistics:
Poisson
Scan For Areas With:
High Rates
Low Rates
High or Low Rates
Time Aggregation
Units: Year
Month
Day
Length: 1 Years
Monte Carlo Replications (0, 9, 999, or value ending in 999): 999
Advanced >>


F:\ENSEIGNEMENT\COURS\master EISIS\2008-2009\OPT17\opt17_data1
SaTScan v8.0
Program run on: Tue Mar 24 10:15:04 2009
Purely Spatial analysis
scanning for clusters with high rates
using the Discrete Poisson model.
SUMMARY OF DATA
Study period.....: 2000/1/1 - 2000/12/31
Number of locations.....: 300
Total population.....: 10524
Total number of cases.....: 869
Annual cases / 100000.....: 8240.2
MOST LIKELY CLUSTER
1 Location IDs included.: 271, 224
Coordinates / radius..: (16.9515,27.9107) / 0.42
Population.....: 45
Number of cases.....: 13
Expected cases.....: 3.72
Annual cases / 100000.: 28829.1
Observed / expected...: 3.50
Relative risk.....: 3.54
Log likelihood ratio..: 7.046426
Monte Carlo rank.....: 119/1000
P-value.....: 0.119
SECONDARY CLUSTERS
Warnings/Errors:
No Warnings or Errors.
Email
Close

livres de référence:

L. Waller, C. Gotway

Applied Spatial Statistics for Public Health Data

eds. Wiley

RS. Bivand, EJ. Pebesma, V. Gomez-Rubio

Applied Spatial Data Analysis with R

eds. Springer, Use R!

jean.gaudart@univmed.fr