Cluster Analysis using SaTScan

Diana Gomez Barroso

Rebeca Ramis Prieto

Outline

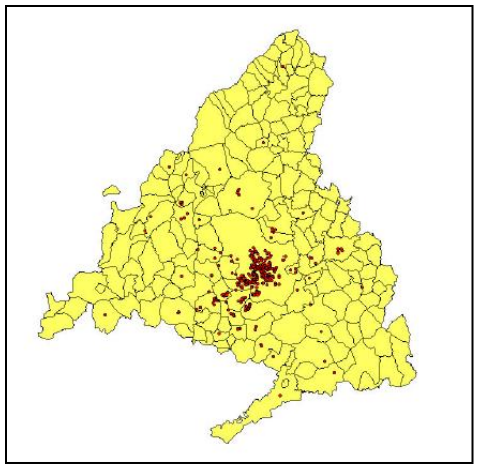
1. Clusters
2. Cluster Detection
3. Spatial and spatio-temporal Scan Statistic
4. Case Study

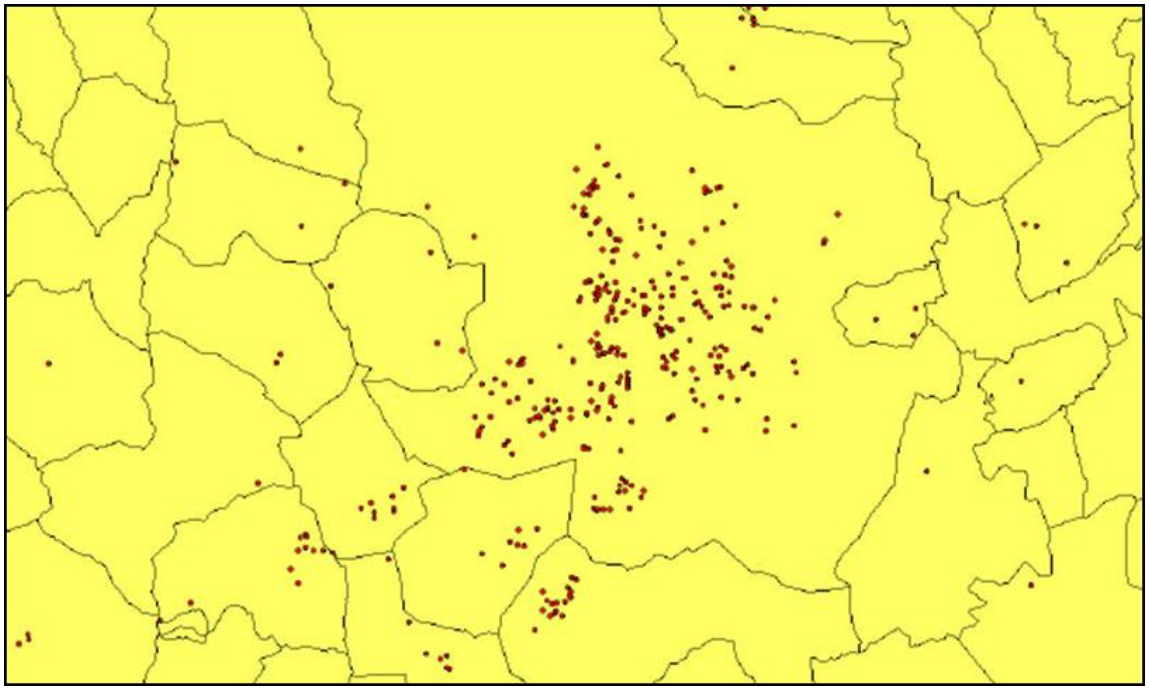
What is a cluster?

What is a cluster?

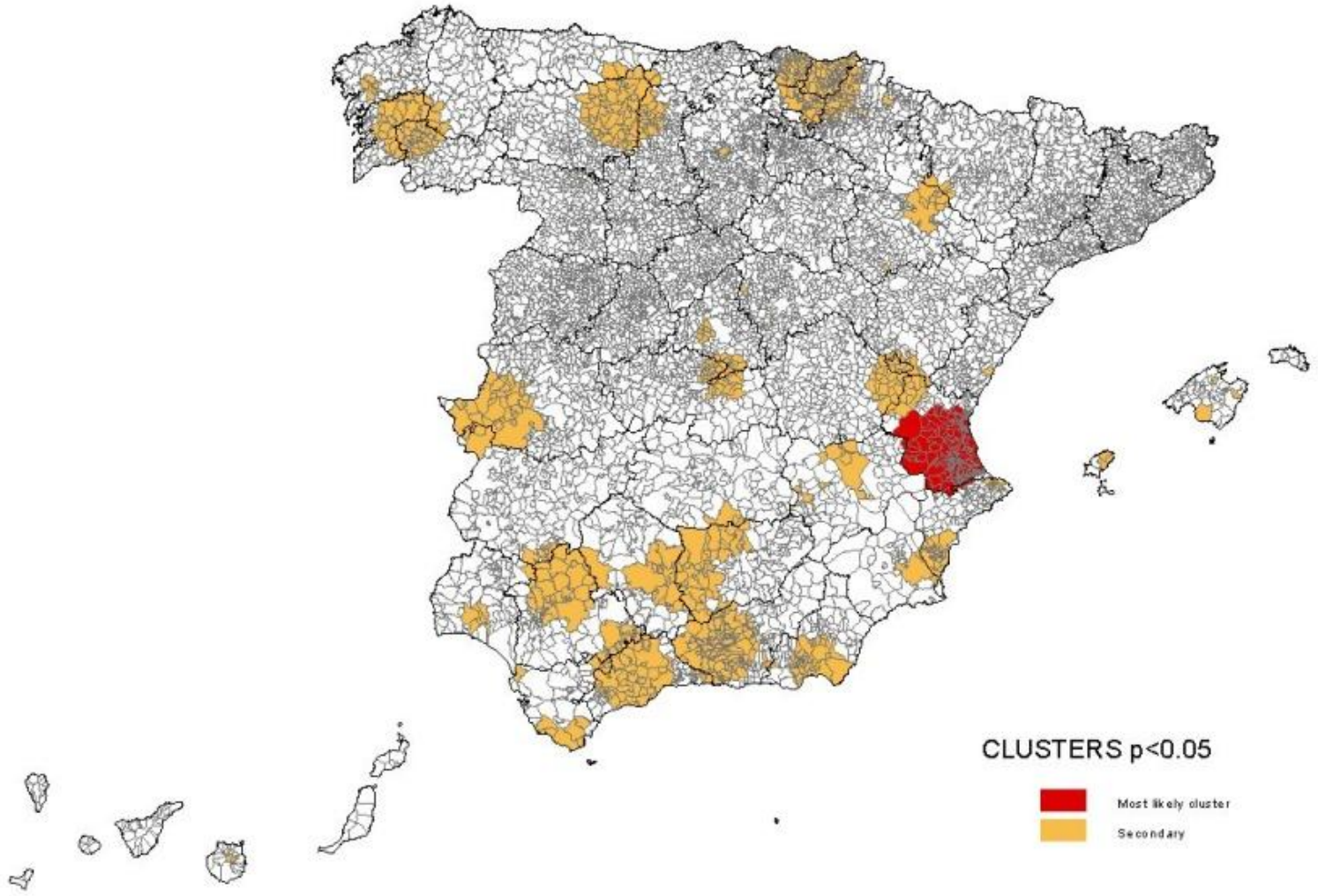
1. An unusual collection of events.
2. Unusual aggregation of real or perceived health events.
3. A geographically and/or temporally bounded collection of occurrences:
    - of a disease already known to occur typically in clusters, or
    - of sufficient size and concentration to be unlikely to have occurred by chance, or
    - related to each other through some social or biological mechanism, or having a common relationship with some other event or circumstance (Knox, 1989)

Point data





Area data (aggregated)



Few issues

- Over what scale does any clustering occur?
- Are clusters merely a result of some obvious *a priori* heterogeneity in the study region?
- Are they associated with proximity to other features of interest, such as transport arteries or possible point sources of pollution?
- Are events that cluster in space also clustered in time?

Outline

1. Clusters
2. Cluster Detection
3. Spatial and spatio-temporal Scan Statistic
4. Case Study

Global and Local Tests

Global tests detect the presence or absence of clustering over the whole study region without specifying the spatial location.

Local tests additionally specify the location and if extended to consider temporal patterns, can specify spatio-temporal clusters.

A special case of local tests is the *focused* test which is used to detect raised incidence of disease around some pre-specified source, such as an pollutant source.

Point Based Methods

Global Measures

Ripley's K Function (and its variants)

Local Measures

Kernel Density Estimation

Kuldorff's Spatial Scan Statistic

Area Based Methods

Global Measures

Moran's *I*

Local Measures

Local Moran's *I*

Kuldorff's Spatial Scan Statistic

Outline

1. Clusters
2. Cluster Detection
3. Spatial and spatio-temporal Scan Statistic

SaTScan

4. Case Study

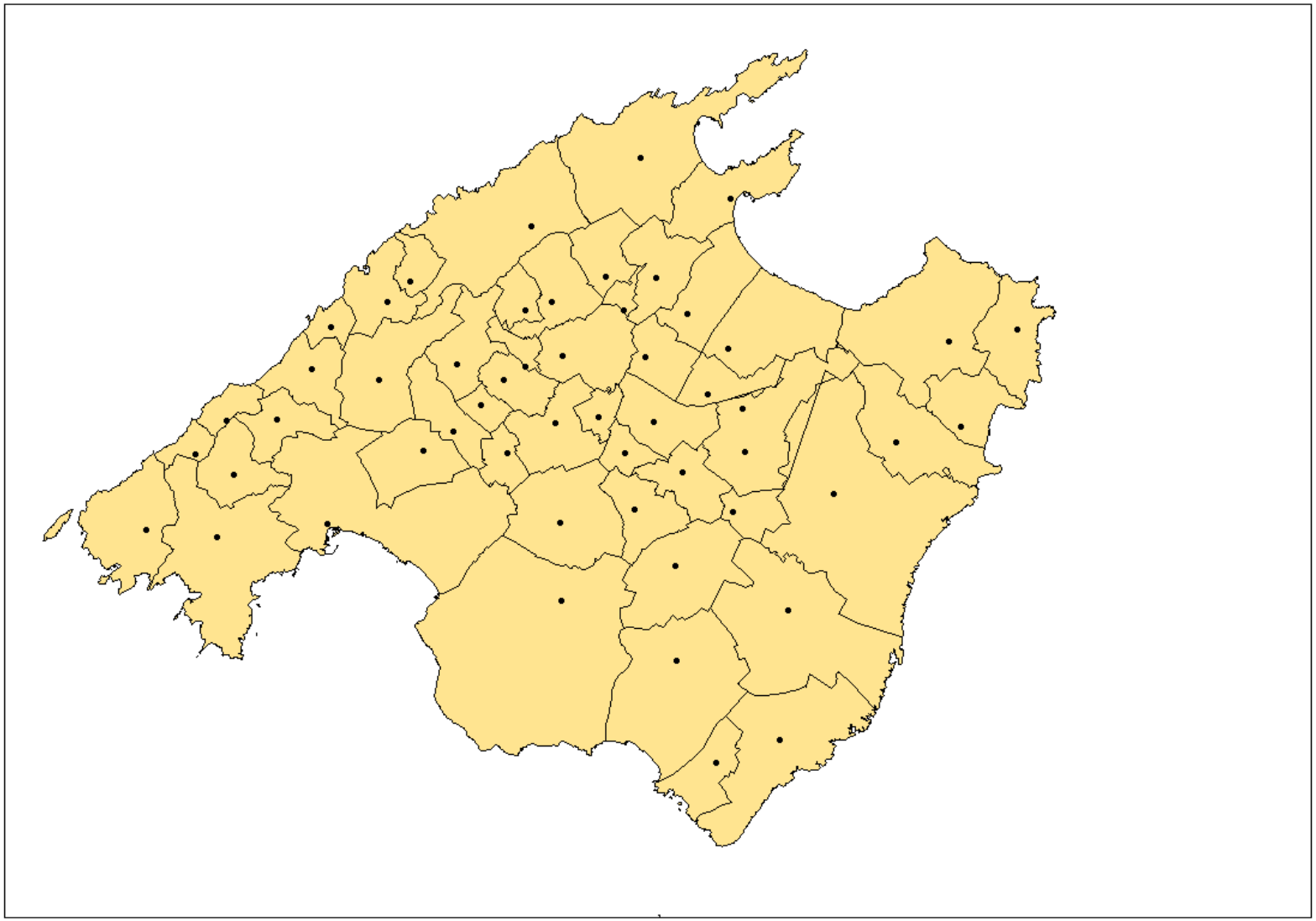
Introduction to SaTScan

- Analyzes spatial, temporal and spatio-temporal data through a scan statistic
- Quick assessment of potential clusters
- Developed by Martin Kuldorff for the National Cancer Institute
- Freely available from www.satscan.org

Objectives

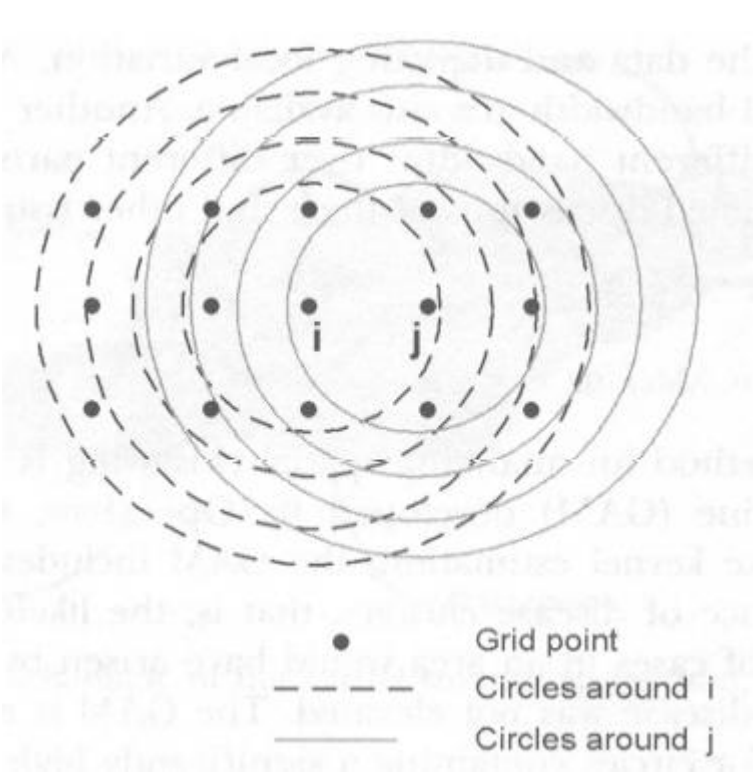
- To perform geographical surveillance of disease, to detect spatial or space-time disease clusters, and to see if they are statistically significant.
- To test whether a disease is randomly distributed over space, over time or over space and time.
- To evaluate reported spatial or space-time disease clusters, to see if they are statistically significant.
- To perform repeated time-periodic disease surveillance for the early detection of disease outbreaks.

Co-ordinates for area data

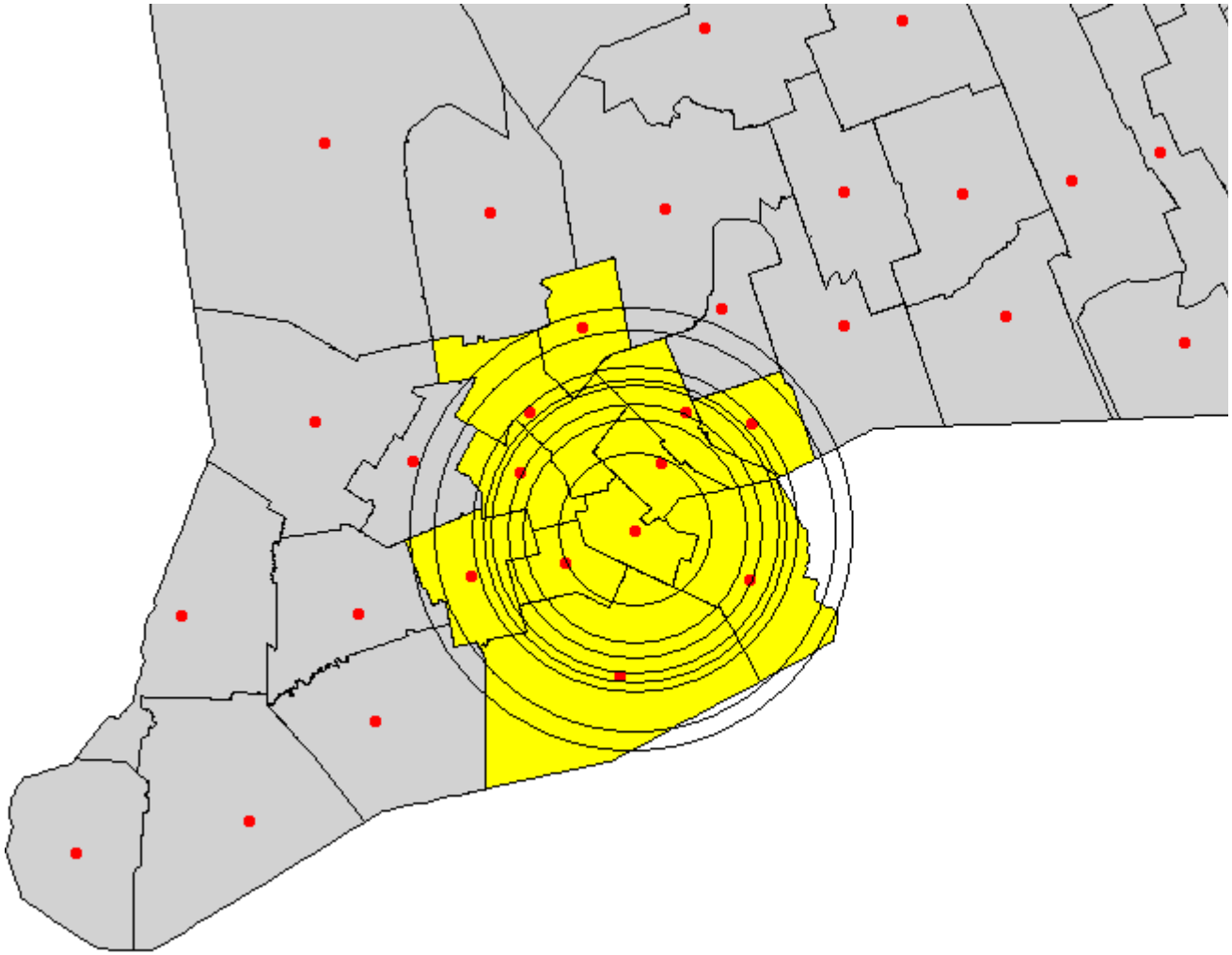


Method

1. A circular scanning window is placed at different coordinates with radii that vary from 0 to some set upper limit.
2. For each location and size of window:
   - H_A = elevated risk within window as compared to outside of window



Method



Method

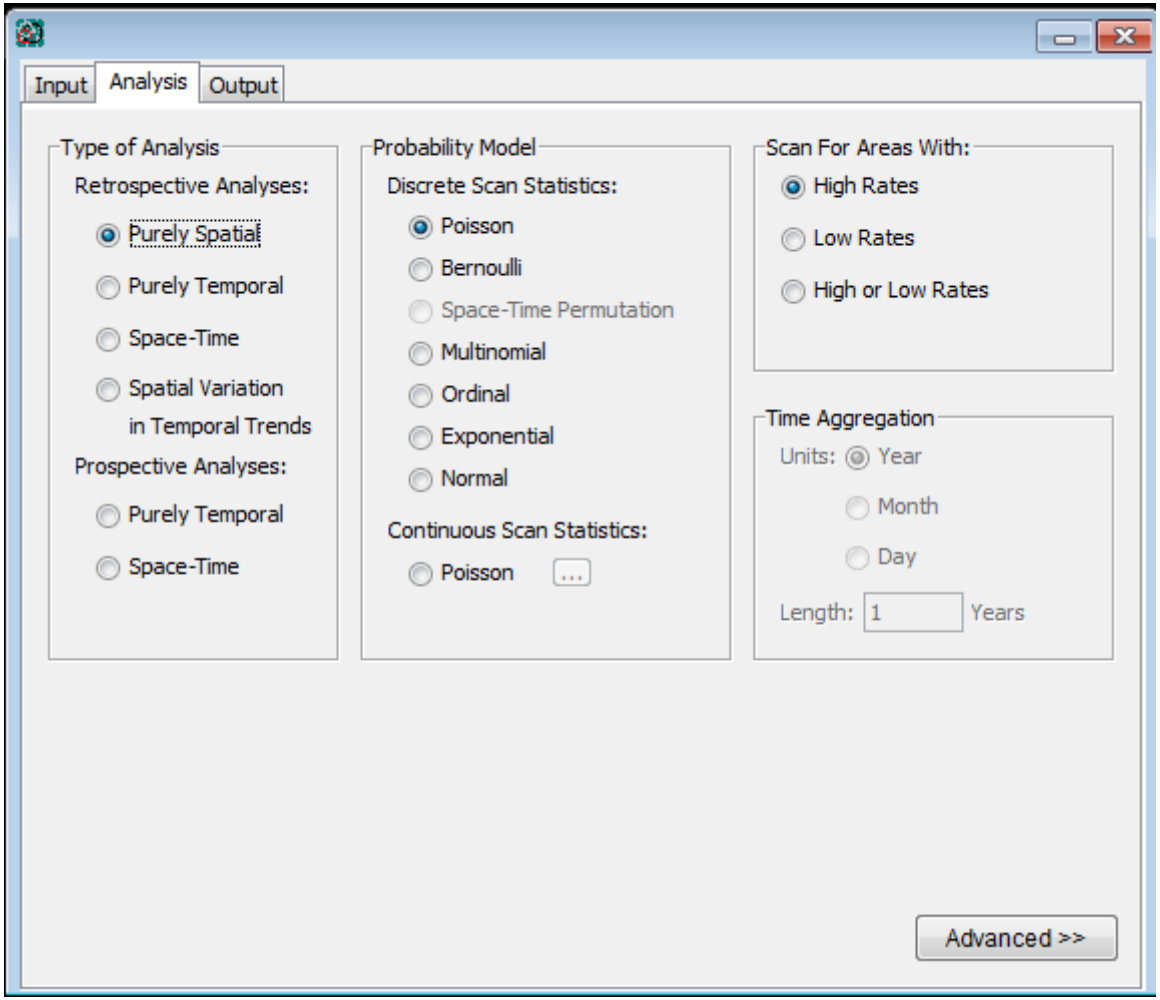
Likelihood function is created depending on model selected (more later…..)

- Likelihood = hypothetical probability that an event that has already occurred would yield a specific outcome
- Function is maximized over all window locations and sizes
  - The one with the maximum likelihood is most likely cluster (least likely to have occurred by chance)
  - Likelihood ratio for this window becomes maximum likelihood ratio test statistic
  - A p-value is obtained for the cluster by Monte Carlo hypothesis testing

Method

- Indicates whether there is clustering
- Shows us where it is (text description)
- Evaluates its statistical significance
- Produces a relative risk for the cluster
  - Which can be low or high risk
- Does not rely on *a priori* knowledge of any cluster

Method


Input
Analysis
Output
Type of Analysis
Retrospective Analyses:
Purely Spatial
Purely Temporal
Space-Time
Spatial Variation in Temporal Trends
Prospective Analyses:
Purely Temporal
Space-Time
Probability Model
Discrete Scan Statistics:
Poisson
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
Advanced >>

Method. Poisson models

Should be used when background population reflects a certain risk mass such as total persons in an area

Expected number of cases is directly proportional to the total number of persons

Example – # cases of cancer per 100,000 persons

Method. Poisson models

Likelihood function:

$$\left(\frac{c}{E[c]}\right)^{c}\left(\frac{C-c}{C-E[c]}\right)^{C-c} I()$$

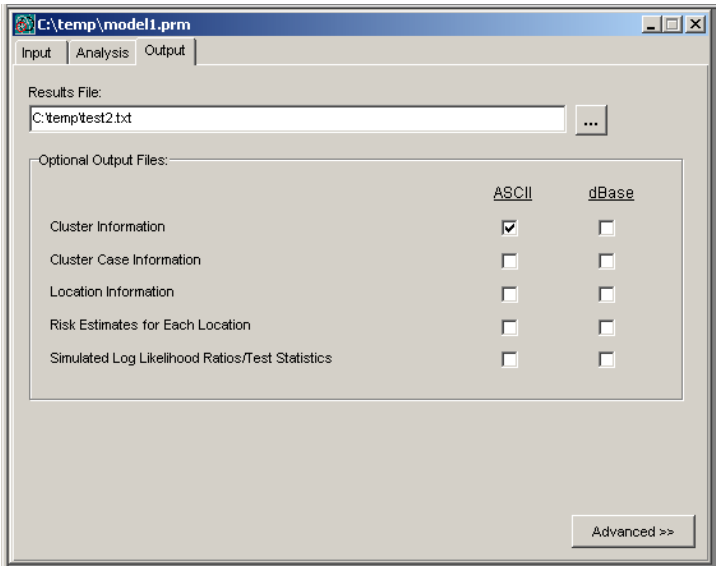
C = total number of cases

c = observed number of cases in window

E[c] = covariate adjusted expected number of cases in window

I() = indicator function

Output. Poisson models



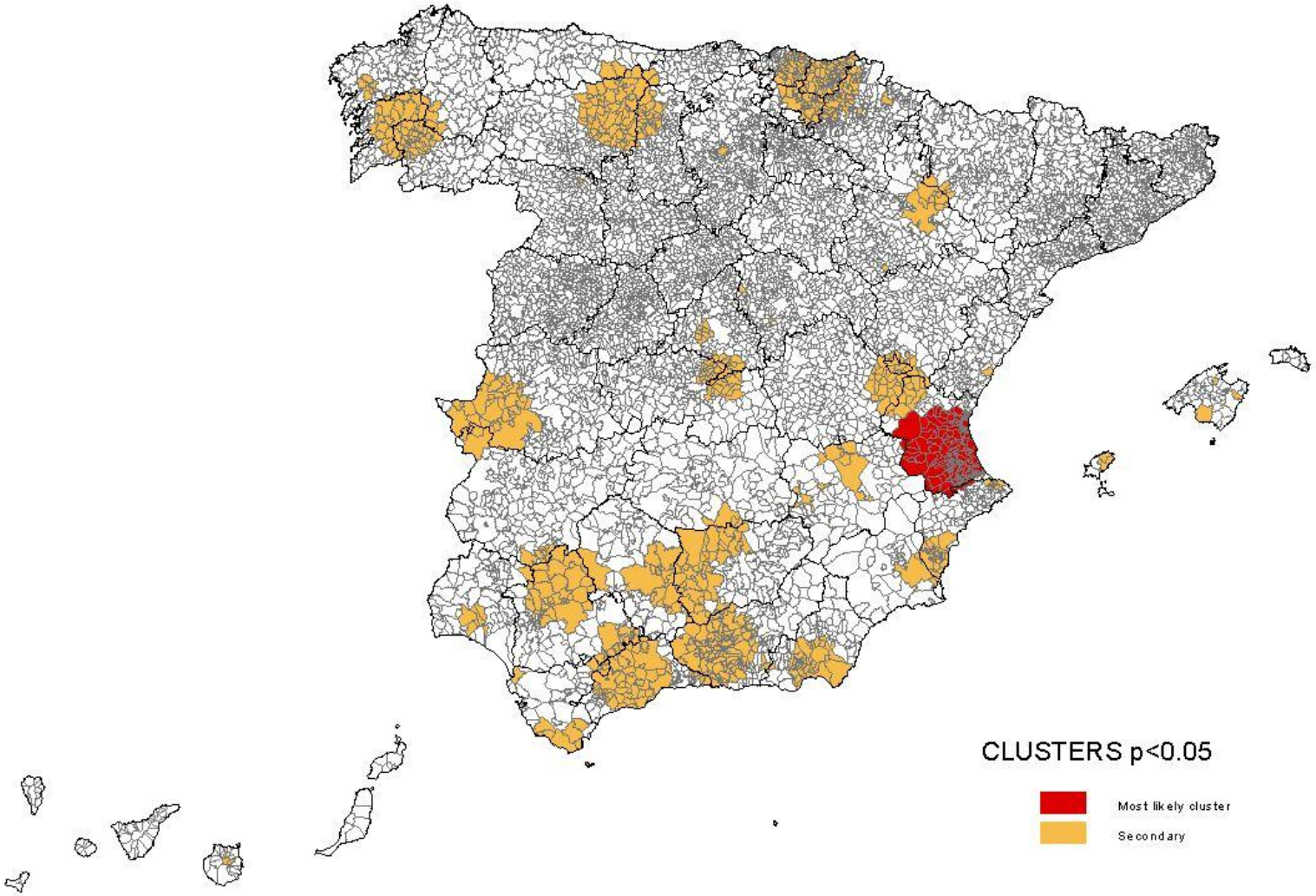
Need to import this into a GIS for visualization

test2.gis.txt - Notepad

File Edit Format Help

ID	CLUSTER	p	Obs	Exp	Rate
35250169	1	0.001	40	3.3	12.117
35250023	2	0.001	40	6.47	6.178
35250140	3	0.001	252	401.99	0.627
35250138	3	0.001	252	401.99	0.627
35250139	3	0.001	252	401.99	0.627
35250168	3	0.001	252	401.99	0.627
35250136	3	0.001	252	401.99	0.627
35250137	3	0.001	252	401.99	0.627
35250134	3	0.001	252	401.99	0.627
35250653	3	0.001	252	401.99	0.627
35250141	3	0.001	252	401.99	0.627
35250133	3	0.001	252	401.99	0.627
35250652	3	0.001	252	401.99	0.627
35250132	3	0.001	252	401.99	0.627
35250655	3	0.001	252	401.99	0.627
35250650	3	0.001	252	401.99	0.627
35250135	3	0.001	252	401.99	0.627
35250651	3	0.001	252	401.99	0.627
35250584	3	0.001	252	401.99	0.627
35250656	3	0.001	252	401.99	0.627
35250648	3	0.001	252	401.99	0.627
35250649	3	0.001	252	401.99	0.627
35250142	3	0.001	252	401.99	0.627
35250654	3	0.001	252	401.99	0.627
35250657	3	0.001	252	401.99	0.627
35250658	3	0.001	252	401.99	0.627
35250647	3	0.001	252	401.99	0.627
35250659	3	0.001	252	401.99	0.627
35250131	3	0.001	252	401.99	0.627
35250640	3	0.001	252	401.99	0.627
35250121	3	0.001	252	401.99	0.627
35250663	3	0.001	252	401.99	0.627
35250638	3	0.001	252	401.99	0.627
35250122	3	0.001	252	401.99	0.627
35250661	3	0.001	252	401.99	0.627
35250585	3	0.001	252	401.99	0.627
35250660	3	0.001	252	401.99	0.627

Output. Poisson models

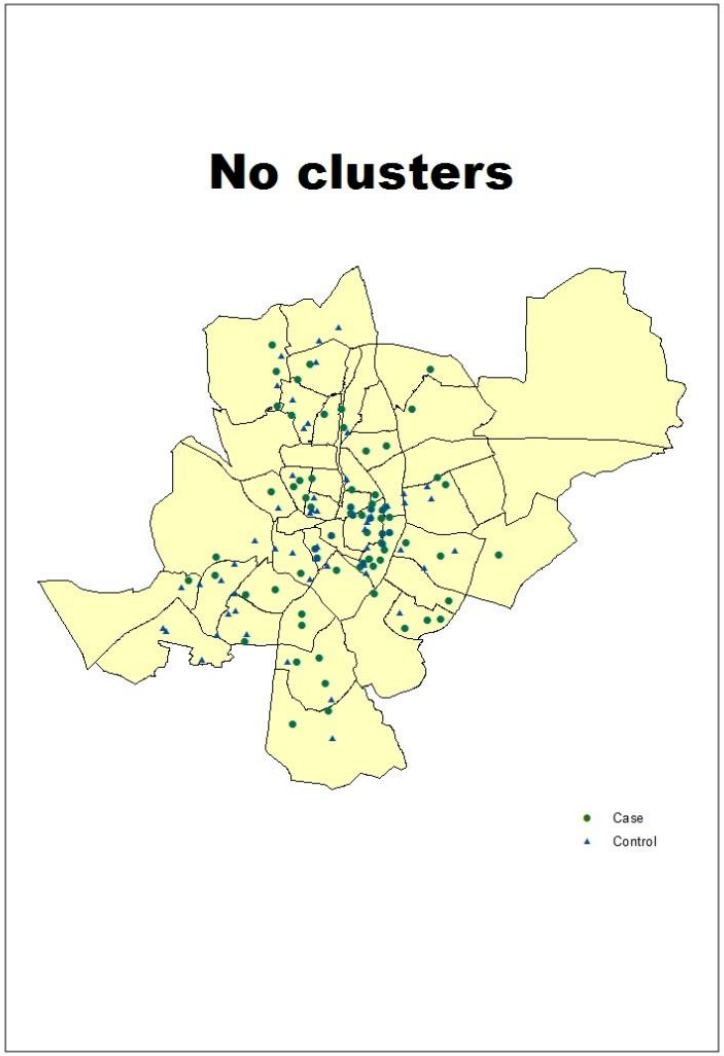


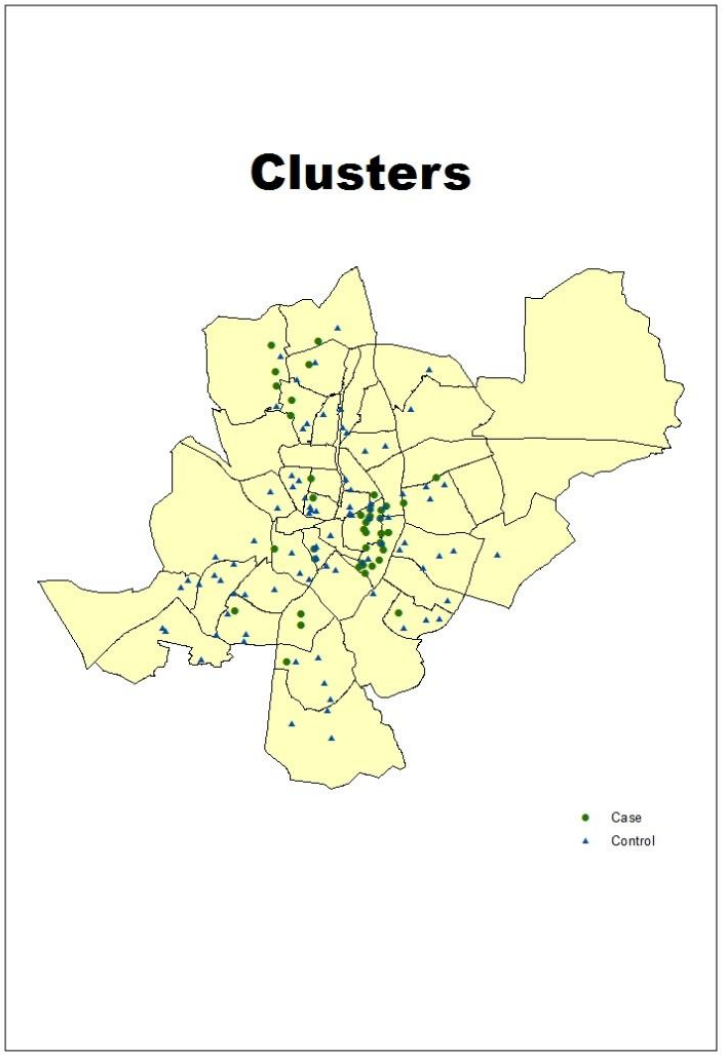
Method. Bernoulli models

Bernoulli process is a discrete-time stochastic process based on *Bernoulli trials*

- An experiment whose outcome is random and can be either of two possible outcomes, “success” and “failure”
- Values expressed as 0 or 1 (non-cases or cases)

Method. Bernoulli models





Method. Bernoulli models

Scan similar to Poisson, visiting each event

Likelihood function:

$$\left(\frac{c}{n}\right)^{c}\left(\frac{n-c}{n}\right)^{n-c}\left(\frac{C-c}{N-n}\right)^{C-c}\left(\frac{(N-n)-(C-c)}{N-n}\right)^{(N-n)-(C-c)} I()$$

C = total number of cases in dataset

c = observed number of cases in window

n = total # of cases and controls in window

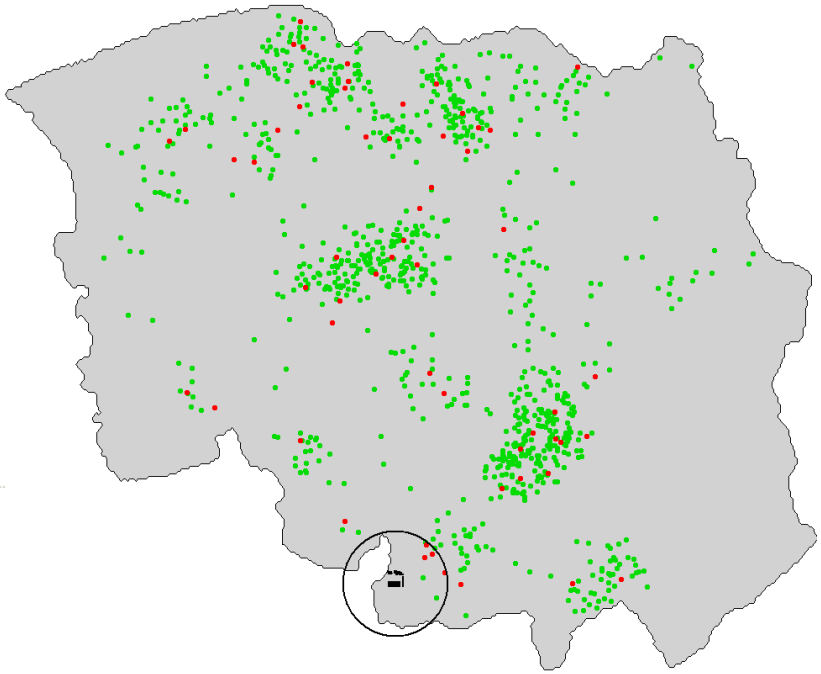
N = total number of cases and controls in dataset

Focused tests

Focused tests can be conducted by providing the coordinates of one or more point sources. In these instances, the scan will only evaluate clusters around these sources and bypass the rest of the coordinates

```
MOST LIKELY CLUSTER

1.Location IDs included.: 220, 1035, 1033, 1034, 75, 9, 1032
  Coordinates / radius..: (354558,413465) / 1353.79
  Population.....: 7
  Number of cases.....: 4
  Expected cases.....: 0.41
  Observed / expected...: 9.764
  Relative risk.....: 10.426
  Log likelihood ratio..: 6.877447
  Monte Carlo rank.....: 6/1000
  P-value.....: 0.006
```



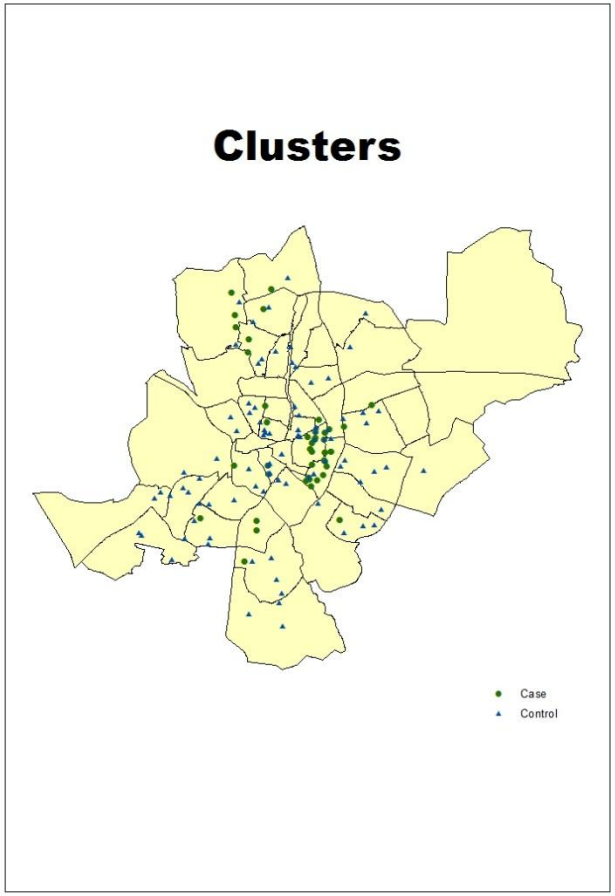
Outline

1. Clusters
2. Cluster Detection
3. Spatial and spatio-temporal Scan Statistic
4. Case Study

Case Study

1. Legionella
2. Leukemia
3. TB
4. Hepatitis A

- Find space and spacio-temporal cluster.
- Give possible causes



Madrid Center