

Further Methods for Point Pattern Analysis

Bailey and Gatrell

Chapter 4

Lecture 12
October 15, 2009

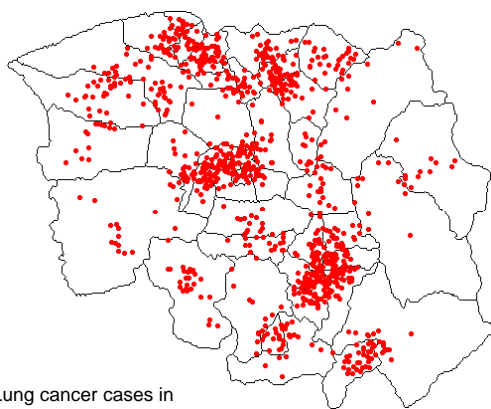
1

Variations in Population

- Certain types of events will exhibit clustering due to heterogeneity in the underlying distribution

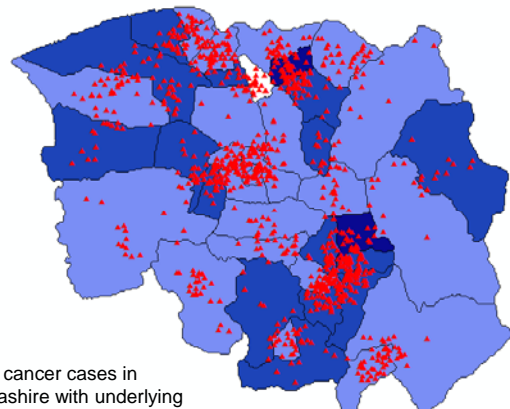
e.g disease cases or crimes will tend to cluster where the population is higher
- In such cases we need to “correct” for the underlying variations in the population
- The more appropriate hypothesis test is against a heterogeneous Poisson process with varying intensity rather than CSR.

2



Lung cancer cases in Lancashire

3



Lung cancer cases in Lancashire with underlying population

4

Use of Kernel Estimates

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^n \frac{1}{\tau^2} k\left(\frac{s-s_i}{\tau}\right)$$

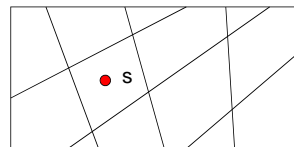
- Kernel estimate estimates events per unit area
- Reconsider as estimate of events per unit population
- Divide estimated event intensity at location s by an estimate of population density at s
- How to estimate population density at s ?

5

Estimations of Population Density

- Can take population count for a census unit and convert to density by dividing by the census unit area
- For the population density at s , use the population density for the census unit in which s falls

$$pd_j = \frac{P_j}{A_j} \quad \begin{array}{l} \text{Population for unit } j \\ \text{Area for unit } j \end{array}$$



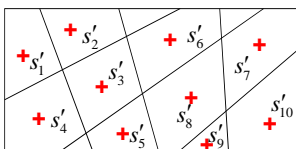
$$\hat{\rho}_\tau(s) = \frac{\sum_{i=1}^n \frac{1}{\tau^2} k\left(\frac{s-s_i}{\tau}\right)}{pd_j(s)}$$

6

Estimations of Population Density

$$s'_j \quad j = 1, \dots, m$$

Locations where populations are recorded



$$y_j \quad j = 1, \dots, m$$

Population values at s'_j

$$\hat{\lambda}'_\tau(s) = \sum_{j=1}^m \frac{1}{\tau^2} k\left(\frac{s-s'_j}{\tau}\right) y_j \quad \begin{array}{l} \text{Estimate of population per} \\ \text{unit area} \end{array}$$

7

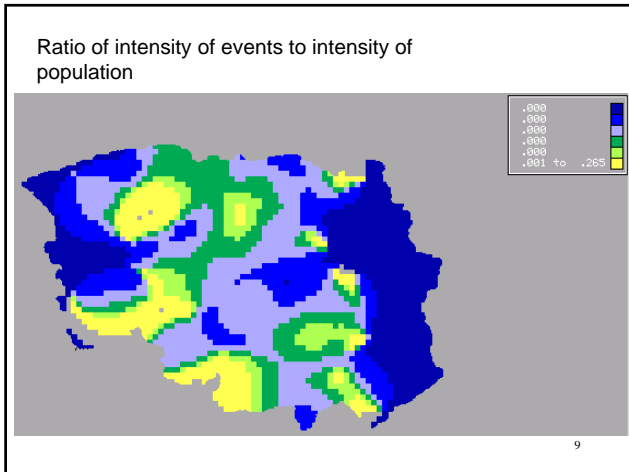
Estimations of Events per Unit Population

$$\hat{\rho}_\tau(s) = \frac{\sum_{i=1}^n k\left(\frac{s-s_i}{\tau}\right)}{\sum_{j=1}^m k\left(\frac{s-s'_j}{\tau}\right) y_j}$$

Ratio of kernel estimates: Divide estimate of events per unit area by population per unit area

- Generally use same kernel and bandwidth for each estimate, but not necessary
- Good estimates of either alone may not lead to good estimates of the ratio
- Small changes in the estimates of population density in regions where value is low lead to large variations in ratio

8



Estimations of Events per Unit Population

- Can apply similar approach to a set of "controls" considered representative of the population variation

Assume m events occurring at points S'_j $j = 1, \dots, m$

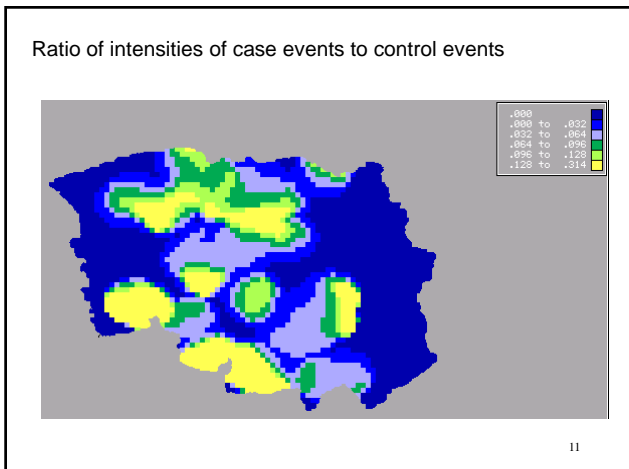
$$\hat{\rho}_\tau(s) = \frac{\sum_{i=1}^n k\left(\frac{s - S_i}{\tau}\right)}{\sum_{j=1}^m k\left(\frac{s - S'_j}{\tau}\right)}$$

Kernel estimate of the cases process

Kernel estimate of the control process

Assume same kernel and bandwidth

10



Interaction of Cases and Controls

- Assume a set of n_1 cases of a disease in region R
- Also assume a random sample of a set of locations in R of n_2 controls from the same population at risk
- Yields $n=n_1+n_2$ events in R of two types: cases and controls
- Can test for clustering of cases relative to controls
- Are cases the same as a random sample from the set of cases plus controls?
- Hypothesis of *random labeling*

12

Hypothesis of Random Labeling

- Random labeling neither implies nor is implied by independence
- Only equivalent if both are CSR processes
- Can use K function to test hypothesis of random labeling
- Under random labeling the pattern of either cases or controls represents a random thinning of the combined point pattern
- K functions are invariant under random thinning

Theoretical expectation is thus

$$K_{11}(h) = K_{22}(h) = K_{12}(h)$$

cases controls combined

13

Hypothesis of Random Labeling

Assume cases are type 1 events and controls are type 2 events

Plot $K_{11}(h) - K_{22}(h)$ against h
to check for departures from random labeling

Peaks represent clustering of cases over controls

To assess significance

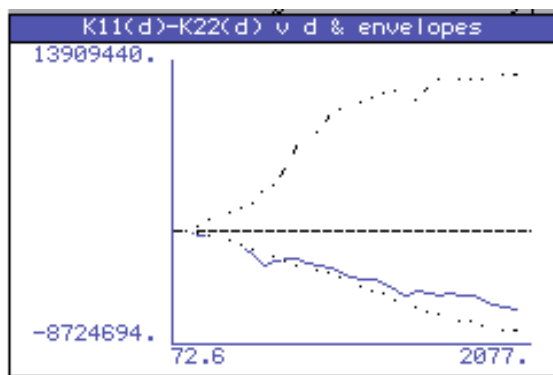
Upper and lower simulation envelopes can be developed from estimates of $\hat{K}_{11}(h)$ and $\hat{K}_{22}(h)$ in repeated simulations

For the simulations:

use the n_1+n_2 locations but randomly reassign case labels to n_1 of the locations.

14

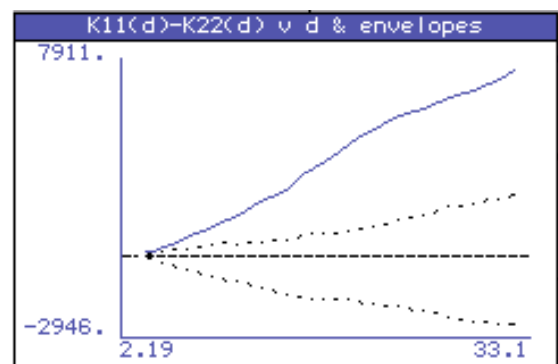
Test of random labeling hypothesis for larynx versus lung cancers



Indicates larynx slightly more dispersed

15

Tests of random labeling between black (K_{11}) and white (K_{22}) crimes



16

Clustering around a Point Source

Statistical problem: estimate the risk of disease among people living near a possible point source of environmental pollution and compare with the risk in the general population

Weaknesses in statistical approaches

- Standard rates may not apply around the source
- A hypothesis about the source may have been formed and some boundaries chosen based on informal knowledge about number of cases
- Pre-selecting an area around the source on which to base an estimate of risk is difficult when the scale of the possible effect is unknown

17

Clustering around a Point Source

- test hypothesis that risk of the disease in Area A is greater than it would have been in the absence of the source

O= observed mortality

E= expected number of deaths based on standard mortality rates

O/E is estimated relative risk

θ is true relative risk

Random variability in O is Poisson distributed with mean θE

$H_0: \theta=1$

In the absence of the source the standard rates apply

18

Clustering around a Point Source

Can use ratio of kernel estimates (cases/ controls) as an exploratory method

Modeling Approach

Use a heterogeneous Poisson model where intensity of case events $\lambda(s)$ is expressed as multiplicative function of the intensity of the background population (controls) and distance from a suspected source

$$\lambda(s) = \rho \lambda'(s) f(h; \theta)$$

ρ is ratio of number of cases to controls

$\lambda'(s)$ is background population intensity

$f(\cdot)$ is a distance decay function

h is distance from an arbitrary point to the source

19

Clustering around a Point Source

Consider different forms for $f(h; \theta)$

$$f(h; \theta) = 1 + \theta_1 e^{-\theta_2 h^2}$$

$\lambda'(s)$ is estimated using kernel estimate

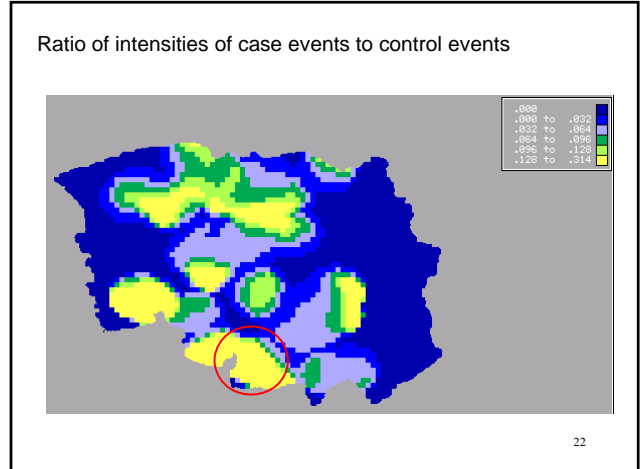
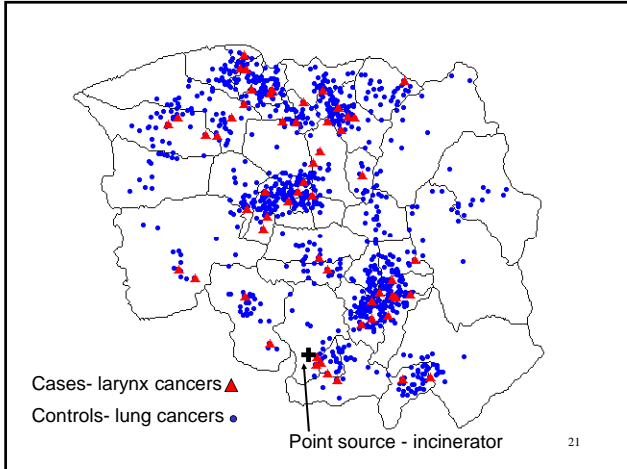
θ_1 and θ_2 are parameters to be estimated

H_0 : There is no distance decay effect or θ_2 equal to zero

Is independent of distance – intensity of controls scaled by a constant

Other covariates can be included in $f(\cdot)$

20



Model Fit

Fit using maximum likelihood estimation; the significance of the model parameters indicates potential increased risk.

Parameters of the fitted model:

$$\theta_1 = 33.74344 \quad \theta_2 = 1.10214 \quad \rho = 0.05532$$

Model Fit: Nominal p-value = 0.013215

The probability p applies to the test of the null-hypothesis $\theta_1 = \theta_2 = 0$. Interpretation: the chance of getting a more extreme outcome for the model parameters is $\leq .013$ (generally considered significant). Thus the model with two parameters is significant, and it appears that there is an elevation in the local risk about the toxic site.

23

Cluster detection tests

Scan statistics are used to detect and evaluate clusters in a temporal, spatial, or space-time setting

They work by gradually scanning a window across time, space or space-time and noting the number of observed and expected observations inside the window at each scan location

The scanning window can be either an interval (in time), a circle or an ellipse in space, or a cylinder with a circular or elliptical base in space-time

Multiple different window sizes are used, each window is a possible candidate cluster.

By searching for clusters without specifying their size or location the method avoids the problem of pre-selection bias.

24

Space-Time Scan Statistic

Cases are assumed to be Poisson distributed with constant risk over space and time under the null hypothesis

Alternative hypothesis specifies different risk inside and outside the scan window.

H_0 : The null spatial model is any inhomogeneous Poisson or Bernoulli process whose intensity (e.g. Poisson parameter) is proportional to some known function, such as population size and risk.

25

Space-time Scan Statistic

Test Statistic:

Scan window is defined and moved systematically throughout the geographic and temporal space of interest. The window size is varied.

A likelihood ratio test statistic over all possible scan windows is calculated, conditioning on the observed total number of cases (C).

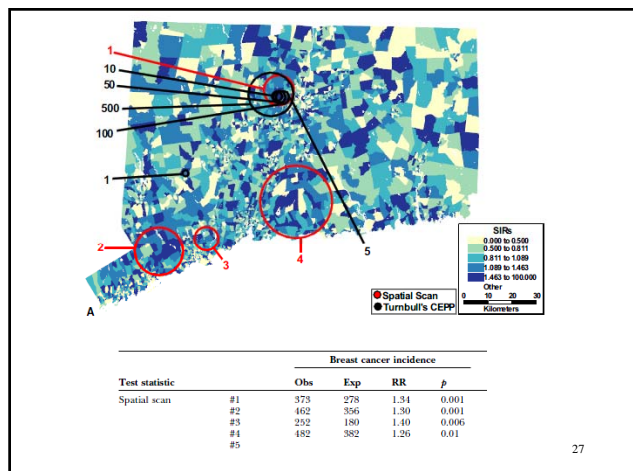
$$\frac{L_{(i,j)}}{L_0} = \left(\frac{c}{E[c]} \right)^c \left(\frac{C-c}{C-E[c]} \right)^{C-c}$$

c = observed cases within the window
 C = total # of cases

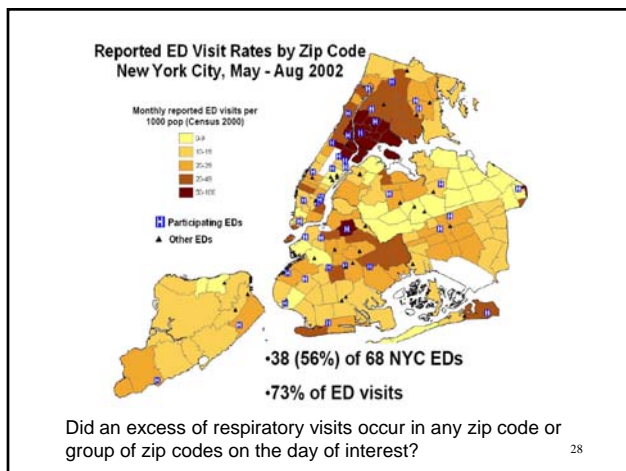
For every potential cluster, calculate likelihood ratio statistic for null hypothesis of a common rate in all cells against an alternative 2 rates: a rate for cells outside cluster and a higher rate for cells inside.

The window for which the likelihood ratio is maximized identifies the most likely cluster (MLC)

26



27



28

Determining the appropriate population at risk

For cancer epidemiology

Numerator = cancer cases

Denominator = census population

For infectious disease surveillance

Numerator = Cases (Respiratory ED visits)

Denominator = ???

29

Calculation of zip code denominator

ED patients in zip code
during 14-day baseline

$$\frac{\text{respiratory: } 80}{\text{total visits: } 1300} = 0.07$$

ED patients in zip
code yesterday

total visits : 90

1-day
lag

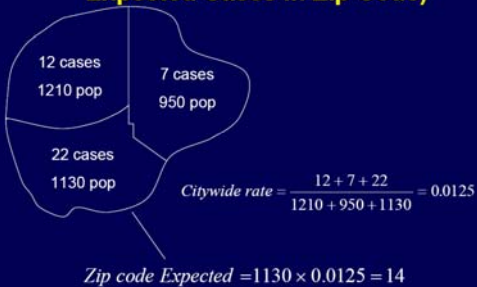
$$\text{Denominator 'pop'} = 0.07 * 90 * 1000 = 6,300$$

*Mostashari and Kulldorff



30

How SatScan Calculates Expected Cases in Zip Code



Space-time Scan Statistic

Identifies a significant excess of cases within a moving window (circular, cylindrical, elliptical)

Provides a measure of how unlikely it would be to encounter the observed excess of cases.

Useful screening tool for evaluating which clusters merit further investigation and which clusters are probably chance occurrences

Kulldorff, M. 1999. Spatial scan statistics: models, calculations, and applications, in *Scan Statistics and Applications*, Glaz, J & Balakrishnan (eds.), Birkhauser, Boston, pp.303-322

<http://www.satscan.org/>

32

Scan limitations

Unusually shaped clusters that may not be detected by the circular or elliptical spatial scan window

Recently, spatial scan statistics for irregular shaped clusters have been proposed, use the same likelihood ratio test formulation.

33

Rash and respiratory data during August 1–30, 2005, geographically aggregated to ZIP codes

