Correcting for Spatial Variations in the Population at Risk Using S-Plus and SPLANCS

STAT 5810 - PROJECT 1

Group 2

Mingyo Chung, Katya Saraeva

Guy Serbin, Vasile Turcu

Tong Yin, Jinsheng You

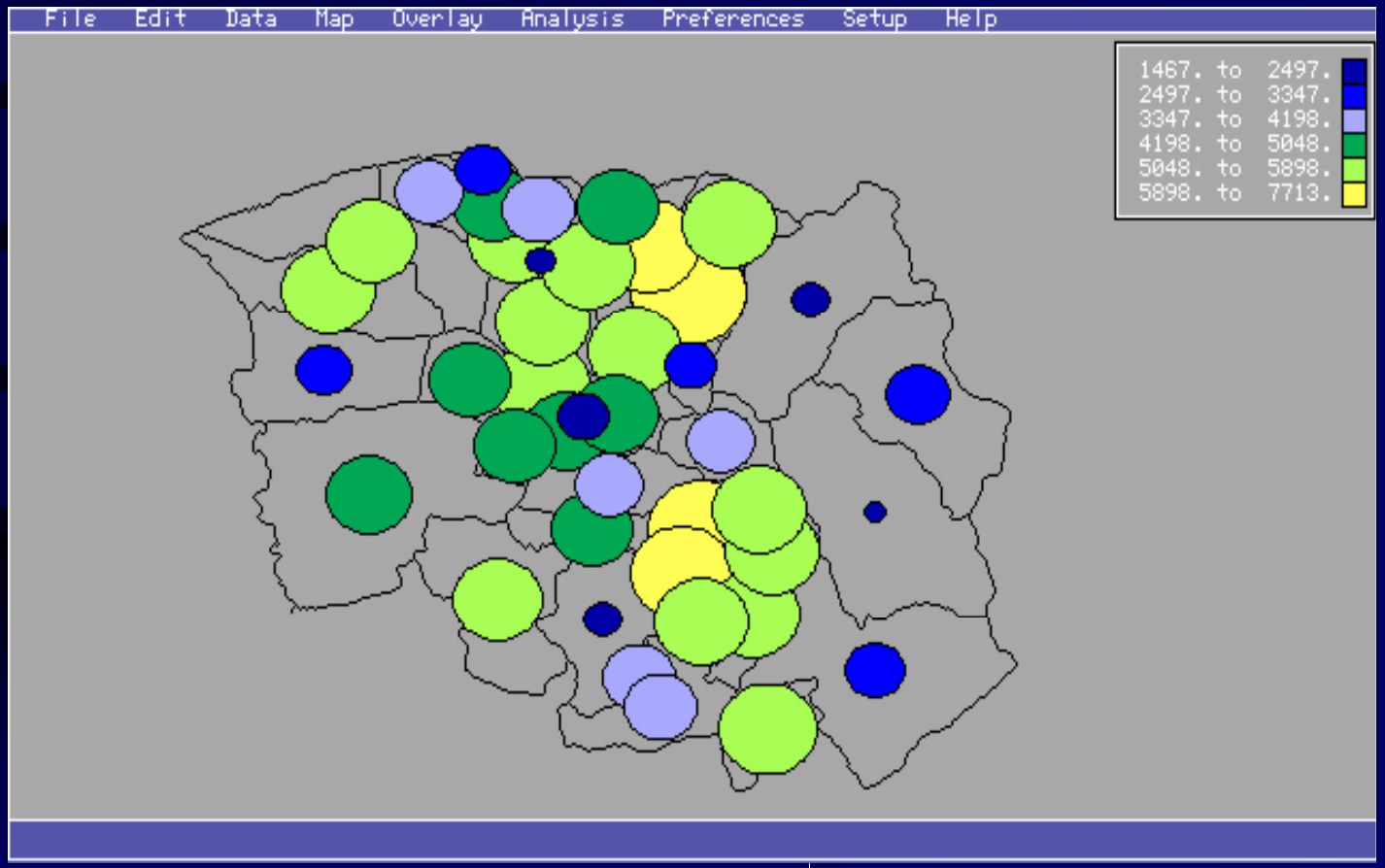
Background

- In the epidemiological cases, the occurrence of diseases is expected to vary with the population density.
- When natural spatial variation in background population exists, instead of comparing the disease occurrence with a CSR process, we test the clustering hypothesis against a heterogeneous Poisson process with varying intensity $\lambda(s)$, e.g, background population, another type of events within the same area.

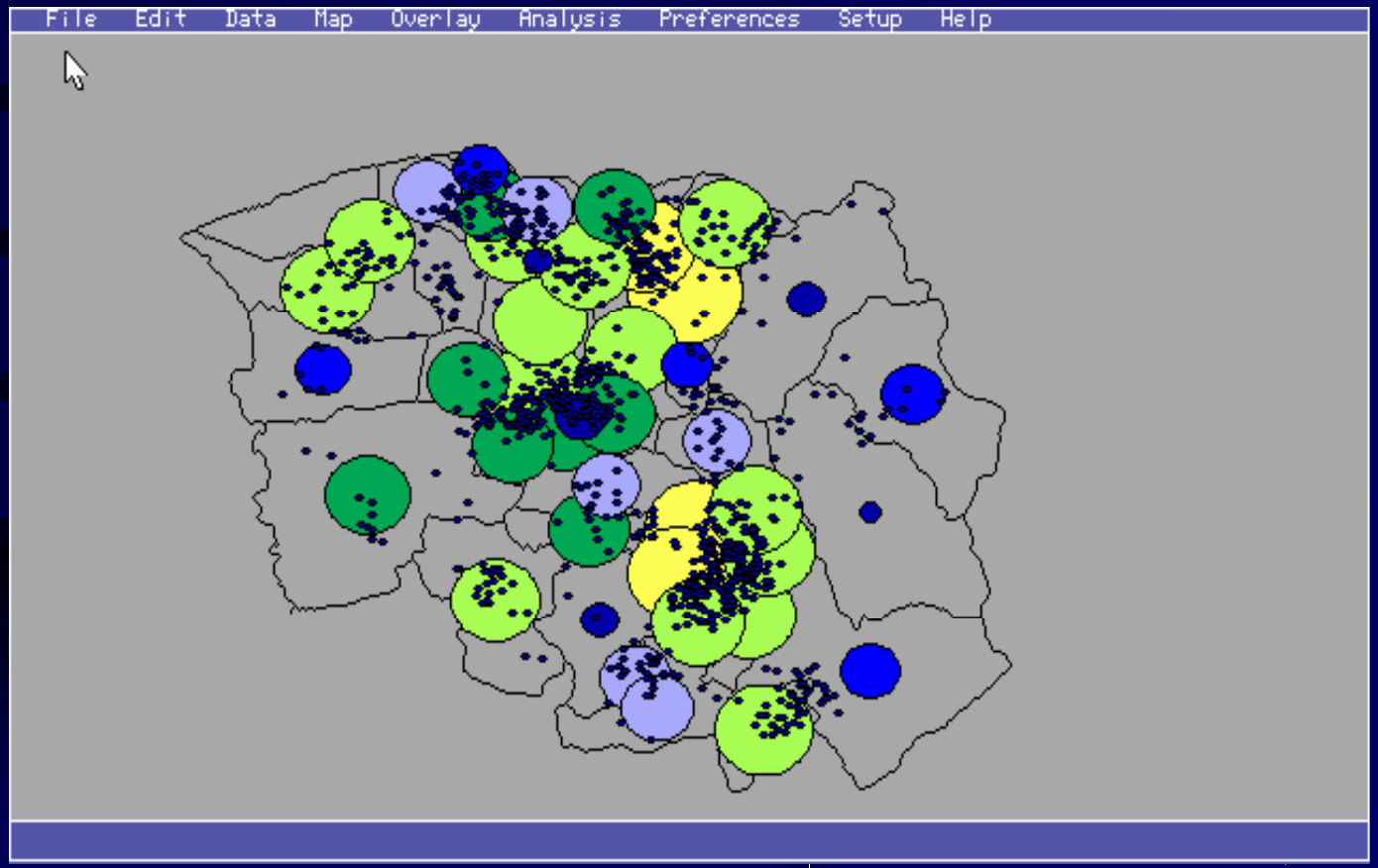
Data

- Data set: larynx and lung cancers of Lancashire in Britain
- The data set consists of five columns of numbers: easting, northing (define the locations of events), Population (expressed as number of people), Lung cancer, and Larynx cancer (1.00 represents occurrence, -999.00 means no data or non-occurrence) of Lancashire.
- Number of cases of lung cancer: 917
- Number of cases of larynx cancer: 57 (from 1974-83)

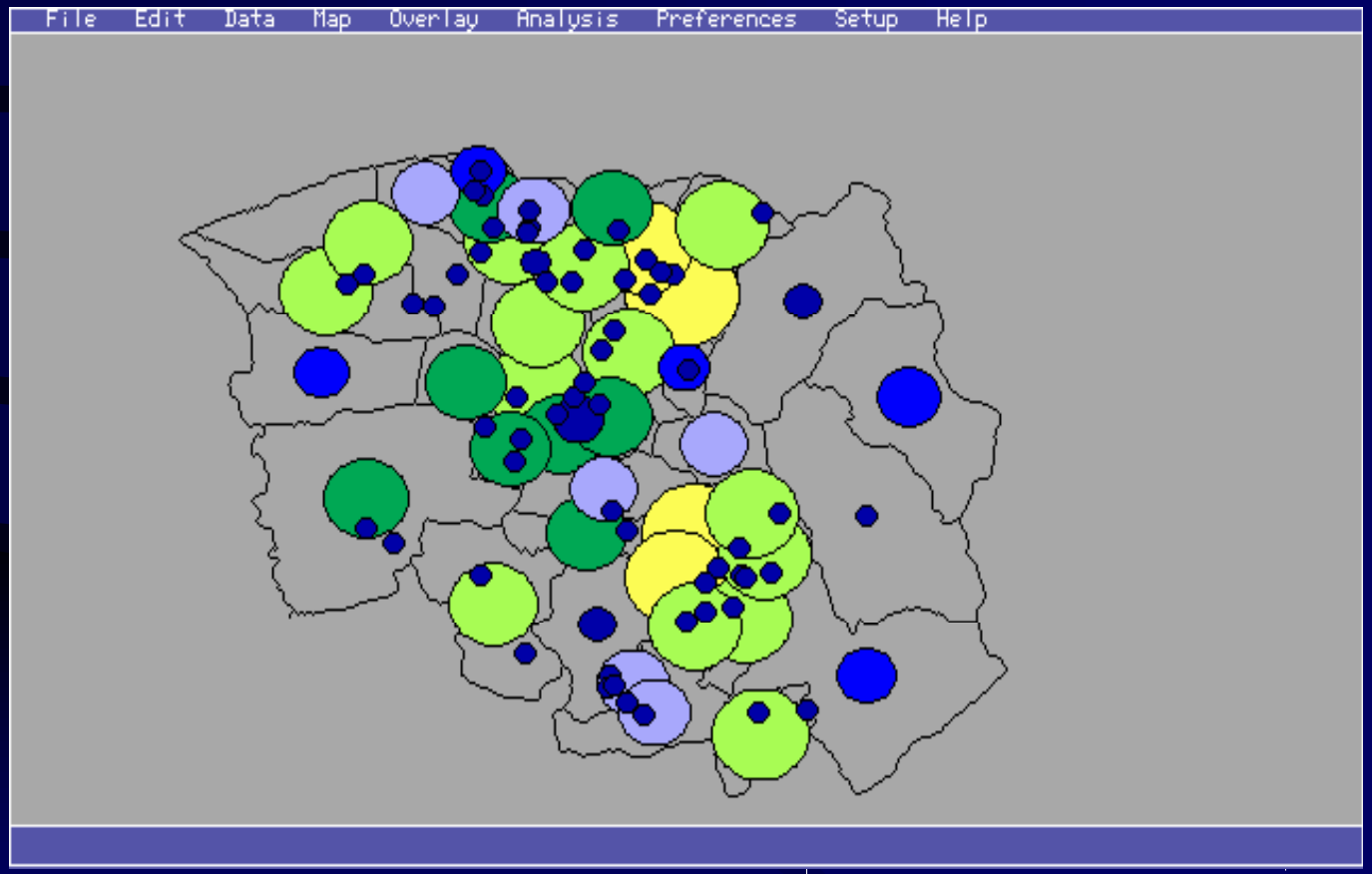
Symbol Map of Population Distribution



Superimposed Dot Map of Lung Cancers on Population Map



Superimposed Symbol Map of Larynx Cancers on Population Map



Using “Cases” and “Controls”

- One task is to test whether the larynx cancers show any clustering relative to the lung cancers.
- A ‘control’ process is used as a surrogate to ‘mimic’ the variations in population at risk, in this case, lung cancer events are the ‘*controls*’.
- The larynx cancers are the ‘*cases*’.
- ‘*Cases*’ is tested against ‘*controls*’.

Random Labeling Hypothesis

- We have n_1 number of '*cases*', n_2 number of '*controls*' within a study region *R*. Then $n=n_1+n_2$ is the total number of two types of events in *R*, which are *'cases' and 'controls'*.
- If there is no clustering of *'cases'* relative to *'controls'*, then the *'cases'* is just a random sample from the pattern of both cases and controls.
- The hypothesis now becomes: random 'labeling' of cases and controls (the marking of events is independent of their locations and is a uniform distribution over the number of types of events)

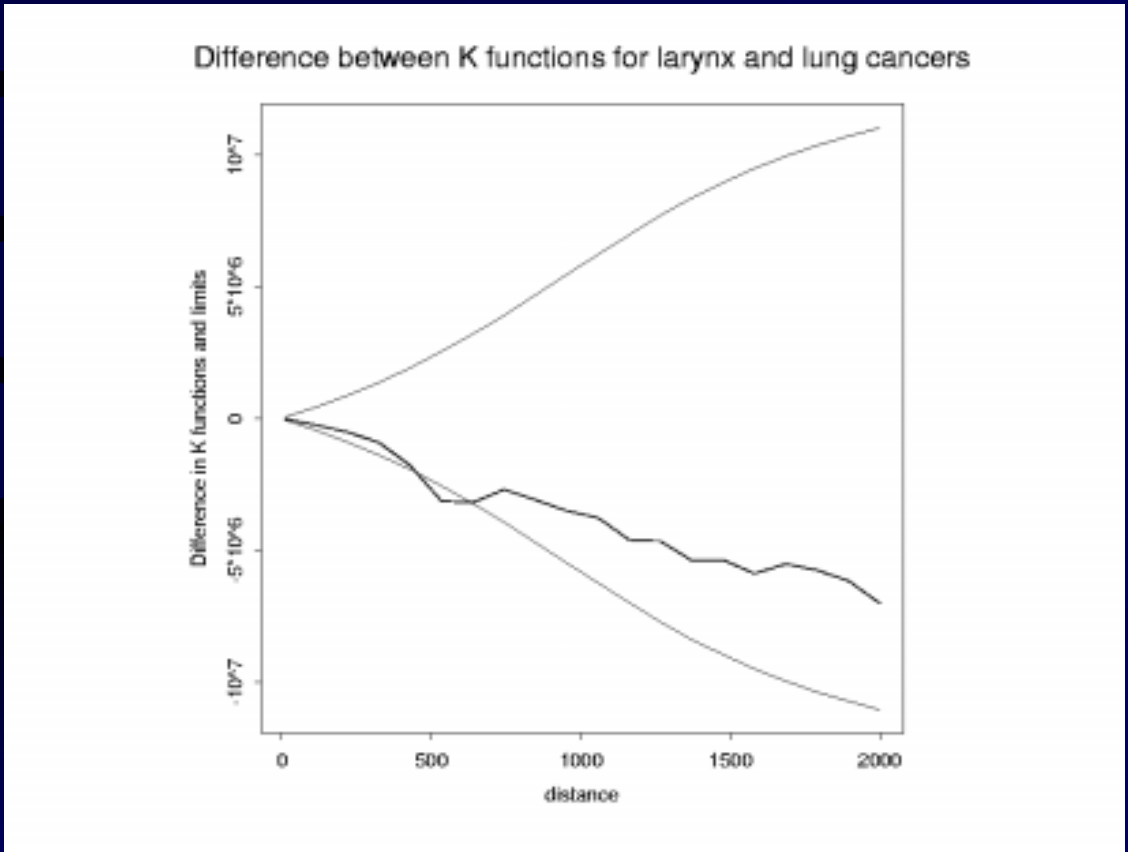
Using K functions

- K functions is a measure of the *reduced second moment* of the observed process.
- We use *K* functions to examine the random 'labeling' hypothesis.
- Under random 'labeling' the pattern of either the *'cases'* or the *'controls'* taken separately represents random 'thinning' of the combined spatial point process.
- K functions are invariant under random 'thinning', it follows that under random 'labeling', we have, $K_{11}(h) = K_{22}(h) = K_{12}(h)$

Plotting

- Therefore, the plot of $\hat{K}_{11}(h)$ - $\hat{K}_{22}(h)$ against h tells if there is departure from random labeling.
- Positive peaks represent spatial clustering of cases over and above the natural environmental spatial clustering of controls.
- Upper and lower simulation envelops for assessing the significance of the peaks are generated in repeated simulation using the fixed n_1+n_2 locations but randomly assigning '*case*' labels to n_1 of these locations.

Plot of difference between K functions for larynx and lung cancers



- The plot shows that the larynx cancers are slightly more dispersed than the lung cancers.