Detection of Spatial Clustering with Average Likelihood Ratio Test Statistics

Assoc Prof Chan Hock Peng, Department of Statistics & Applied Probability

The detection of local clustering in spatial point processes is of interest in epidemiological studies, forestry, geological studies, neural imaging and astronomy. A classical application that will be used here as an illustrative example is the identification of potential sources of environmental pollution that have contributed to higher rates of disease cases for residents living in their vicinity.

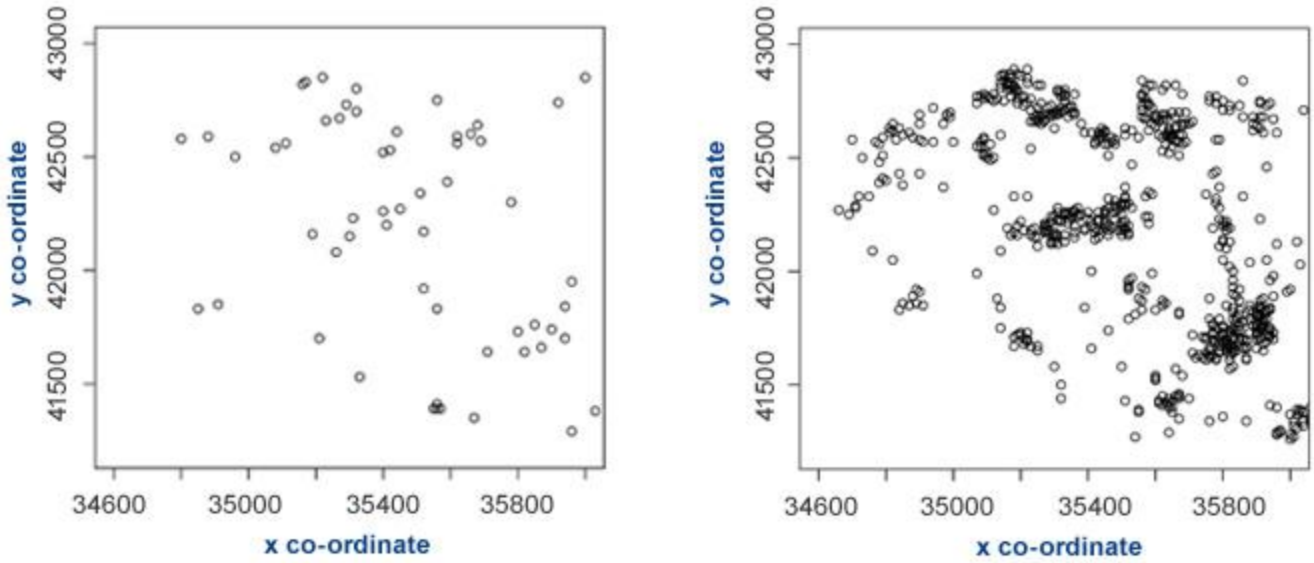

FIGURE 1: Scatter plots of the 58 laryngeal cancer cases (left) and the 978 lung cancer cases (right).

The dataset, see [1], consists of (i) the locations of 58 cases of laryngeal cancer occurring in two districts in Lancashire for the period 1974--85 and (ii) the locations of 978 control cases of lung cancer for the same period and districts. A key feature is a cluster of four laryngeal cancer cases, see the bottom of the left plot of Figure 1, located near an industrial waste incinerator, which is considered a potential source of the cluster of laryngeal cancer cases. We want to test for the presence of local clusters without biasing ourselves a priori with information on the possible sources of the laryngeal cancer cases.

We first construct covering sets which consist of circles of radii 40, 50, 60 and 70, centred at one of the 1036=978+58 points. Hence there are potentially up to 4X1036 covering sets. For each covering set, we compute a generalized likelihood ratio (GLR) score. This score is large when the proportion of laryngeal cancer cases within the covering set is large compared to the proportion of laryngeal cancer cases outside the set. Traditional spatial scan test statistic takes the supremum GLR score over all covering sets. We consider here average likelihood ratio (ALR) test statistics in which we take the average of the GLR scores (not the average of the log GLR score). An important contribution of the paper is that we provide accurate tail probability approximations of the ALR test statistic that allow us to by-pass computer intensive Monte Carlo procedures to estimate p-values. Such computer intensive Monte Carlo procedures are used when computing p-values of spatial scan statistics in the SaTScan software, see [2]. A p-value is an indication of the likelihood of getting the supremum (or average) score by chance alone and often only p-values of less than 0.05 or 0.01 are taken seriously. Monte Carlo simulations are computer experiments to approximate numbers that are hard to compute directly.

An important feature of the tail approximations is that it does not depend on the underlying covariance structure of the dataset and hence can be used when we fit more complicated models that includes additional information, like the age or sex of the subjects. This is especially relevant if we fit case-population datasets rather than the case-control dataset illustrated above.

Publication:

A/Prof Chan Hock Peng - "Detection of spatial clustering with average likelihood ratio test statistics" (***Annals of Statistics***, 2009, 37(6B), 3985-4010)

References:

1. Diggle, P.J., Gatrell, A.C. and Lovett, A.A. (1990). Modelling the prevalence of cancer of the larynx in part of Lanchashire: a new methodology for spatial epidemiology. Spatial Epidemiology, Pion, London.

2. Kulldorff, M. and Information Management Services Inc. (2006). SaTScan User Guide, http://www.satscan.org/techdoc.html.