Change-point Problems
IMS Lecture Notes – Monograph Series (Volume 23, 1994)

DETECTING CLUSTERS IN DISEASE INCIDENCE

BY DANIEL RABINOWITZ
Harvard School of Public Health

This paper is concerned with searching for localized environmental risk factors. The approach taken here uses case-control data to search for clusters of disease cases. In this context, case-control data means a sample of locations associated with diseased subjects (cases) and healthy subjects (controls). A cluster of cases is a region where the number of cases appears to be larger than what would have been expected had the cases occurred randomly in the underlying population. Clusters indicate areas where localized risk factors are likely. The methodology developed here produces a random field over the region where the cases and controls are located. The field is large where there are clusters of cases. Asymptotically, as the number of cases and controls becomes large, the field tends in distribution to a smooth Gaussian field. The operating characteristics of inferential procedures based on the random field may be approximated by considering the random field's limiting distribution.

1. Introduction. Environmental risk factors such as toxic spills, contaminated drinking water and radiation may increase the incidence of cases of birth defects, cancer or disease. When exposure to a risk factor occurs in small areas or during small periods of time, the increased incidence of cases may take the form of a spatial or temporal cluster. If a cluster of cases is detected, health workers can scrutinize the location of the cluster and, one hopes, discover and eliminate localized risk factors that may have caused the increased incidence. Decisions to further investigate locations identified by a cluster detection methodology must be made with reference to the probability of incorrectly detecting a cluster where there is no increased risk.

Case-control data is a sample of locations associated with diseased subjects (cases) and a sample of locations associated with healthy subjects (controls). This paper presents a method for using case-control data to detect clusters of cases. The control data is used to account for non-homogeneities in the density of the population from which the cases arise as suggested in

---

AMS 1991 Subject Classification: Primary 60G70; Secondary 62P99

Key words and phrases: Cluster, point process, screening, scan statistic, tail probability.