Chapter 2

Review of Related Work

In the early part of this chapter, let us first consider some basic statistical concepts and specific terms defined which are preliminary to a further discussion in disease clustering. Next, we would like to review previously proposed tests or methods for the tests of clustering and cluster detection. Finally, we discuss some problems faced with cancer mortality data in Taiwan area based on current proposed methods. Other reviews, especially by Marshall (1991), provide us an overall picture of methods for statistical analysis of spatial patterns of disease. Furthermore, Sankoh and Heiko Becher (2002) offer a flow chart to determine which test is more appropriate for tacking the data in hand.

2.1 Basic Statistical Model

There are commonly two underlying statistical model for most clustering studies: the Bernoulli probability model (e.g. Cuzick and Edwards, 1990; Diggle et al., 1999) and the Poisson probability model (e.g. Whittemore et al., 1987; Openshaw et al., 1988; Besag & Newell, 1991; Turnbull et al., 1998). For some tests, either one of the models can be applied (Kulldorff, 1997). For binary data (as cases or controls) in hand that including finite number of individuals considered, the Bernoulli probability model is commonly applied. Particularly, this model requires point location coordinates for each cases and controls in order to test whether the cases are spatially clustered relative to the controls. On the other hand, under the null hypothesis of the Poisson probability model, number of cases in a region follows a Poisson distribution with a common rate proportional to it population size.

More precisely, considering that the whole study area is divided into K regions, we let $A = \{A_1, A_2, ..., A_K\}$ be the set of census tracts, and $n_k, \xi_k, (x_k, y_k)$ denote the *kth* census tract's observed number of cases, number of individuals at risk or population size, and the coordinate of the population centroid. For the Poisson probability model, under the null hypothesis of no clustering, let N_ks, the number of observed cases, be independent Poisson variables, with the mean of N_k proportional to the number of individuals at risk or population size ξ_k in tract A_k : $E(N_k) = \lambda \xi_k, (k = 1, ..., K)$, where λ is the overall disease or mortality rate that can be estimated from data or empirical results. For the Bernoulli probability model, under the null hypothesis of no clustering, given the total number of cases $n = n_1 + n_2 + ... + n_K$, the tract frequencies are the values of a random sample of size n from a multinomial distribution with parameter $\pi^T = (\pi_1, \pi_2, ..., \pi_K)$, where $\pi_k = E(N_k) \Big/ \sum_k E(N_k) = \xi_k \Big/ \sum_k \xi_k$, for *k*=1,…,*K*.

If we need to adjust for a confounding variable concerned (e.g. age or gender), we shall partition the population into I categories and let n_{ik}, ξ_{ik} denote the observed number of cases and the number of individuals at risk or population size, respectively, in the *ith* category of the confounding factor of the *kth* tract. Thus, for the Poisson probability model, we now assume the cases occur in each of them as independent Poisson process, and the numbers N_{ik} in the *ith* confounding factor group of tract A_k are mutually independent Poisson variables. So, under the null hypothesis, mean of N_{ik} becomes $E(N_{ik}) = \lambda_i \xi_{ik}, (i = 1, ..., I; k = 1, ..., K)$, where ξ_{ik} denotes the number of individuals at risk or population size of the *ith* confounding factor group in tract A_k and, λ_i is the disease or mortality rate of the *ith* confounding factor group that can be estimated from data or empirical results. On the other hand, for the Bernoulli probability model, given the values of $(n_1, n_2, ..., n_I)$, the tract frequencies $(n_{i1}, n_{i2}, ..., n_{iK})$ are mutually independent multinomials with size n_i and the probability vector $\pi_i^T = (\xi_{i1}, ..., \xi_{iK}) \Big/ \sum_k \xi_{ik}, (i = 1, ..., I, k = 1, ..., K)$.

2.2 Tests of Clustering

As pointed out by Marshall (1991), there are two main issues we have to deal with. First, we would like to find out whether there exists a general tendency of clustering to occur and then check out where the clustering area is, if there are any. Second, we would like to answer the question if clusters occur in specific areas.

In most of the literature, the concept of clustering is related to scale of measurement. Sometimes, the public more concern about apparent cluster within a small homogeneous area than within a large scale area because false clusters may be induced by regional differences. So, most importantly, clustering studies are worthwhile if there is a single common mechanism responsible for the cluster. The results of the corresponding tests must be treated with extreme caution because some biases may arise from inaccuracies of case reporting, unknown demographic changes and other reasons.

Generally speaking, three types of tests of clustering are commonly carried out: global tests, local tests and focused test. Many researchers have defined specific definitions for clustering and proposed appropriate tests for detecting them. Most of such test statistics measure similarity or discrepancy between cases or between geographical areas.

2.2.1 Global Tests

Global tests are specially designed to detect clustering that takes place generally in the study region regardless of specific locations.

Some inter-case distance statistics are proposed, for example, the mean Euclidean distance between all pairs of cases by adopting a multinomial model (Mantel and Bailar, 1970; Whittemore, Friend, Brown and Holly, 1987). Grimson et al. (1981) suggested a test statistic that is the count of the number of pairs of labeled objects that are adjacent to one another. The objects can be locations of cases and controls or areas, and adjacency criteria are determined to reflect the kind of clustering under investigation. As pointed out by Besag and Newell (1991), inappropriate test applied may induce apparent clustering even if the underlying rate is constant. The two main reasons are: firstly, more extreme incidence rates or counts

tend to occur among areas with small populations; secondly, for a permutation test, the null hypothesis of the observed incidence rates or counts among the areas are equally probable is invalid for different sized populations.

On the other hand, Cuzick and Edwards (1990) proposed a test statistic that is the sum, over all cases, among the *k* nearest neighbors of each case. If the population density is not uniform, the test statistic of Ross and Davis (1990) that based on the mean Euclidean distance between nearest neighbor cases will avoid the limitation and arbitrary choice of *k*.

2.2.2 Local Tests

Local tests are statistical test for detecting clustering that occurs around specific individual regions within the study area.

Openshaw et al. (1988) performed the analysis of clustering by using Geographical Analysis Machine, GAM, which draw multiple overlapping circles of various sizes, regularly spaced and covering the whole study region. Monte Carlo simulations are used to determine the significance of the number of cases in any circle and various problems arisen have been discussed (Besag and Newell, 1991).

On the other hand, Turnbull (1990) proposed another procedure termed the "Cluster Evaluation Permutation Procedure, CEPP". First, he defined a circular window for each region by including the population for that region and surrounding regions until the total population for the window equals R, the common population size or radius. Then, he measured the maximum number of cases observed in any circle of population radius R as an evidence of clustering based on Monte Carlo simulations' results. If multiple radius values were tested, we may have to use Bonferroni adjustment to account for multiplicity of hypothesis tests. Conversely, Besag and Newell (1991) suggested a method whether a given case is one of the clusters' cases by aggregating the number of nearest neighbor areas, *M*, until there contains at least *k* cases. Again, Monte Carlo simulations are used to provide an exact test of significance. Several different values of k are often necessary to be considered.

Unlike previous methods, to overcome the problem of multiple testing, Kulldorff

and Nagarwalla (1995) proposed a scan statistic to identify significant excesses of cases within a moving circular window, and provided a measure of how unlikely it would be encounter the observed excess in a larger comparison region by applying the likelihood ratio test. The significance level of test is obtained through Monte Carlo hypothesis testing.

2.2.3 Focused Tests

Focused Tests are designed to detect clustering that takes place around a suspected cause for the elevated risk. Many of such tests have been carried out to find the possibility of a linkage between a specific industrial installation (e.g. nuclear plant) and the cases of disease (e.g. childhood leukemia) in the nearby community. When little or nothing is known about the distance scale of any possible effect have led investigator to consider a range of areas about the source. Pre-selection bias may cause the significance of the test becomes meaningless.

Based on distance of cases from a suspected source, Gardner (1989) proposed a method by comparing the observed and expected counts in circles around a source. Similarly, Hills and Alexander (1989) suggested using the comparison of observed and expected mean Euclidean distances from the source as a test statistic. Schulman et al. (1988) proposed a test statistic that is the mean distance from the source after using a map projection to equalize the population density of the areas. Stone (1988), Stone and Bithell (1989) suggested a test for monotonic decay of risk with increasing distance assuming that the levels of exposure to the source are unknown but that the exposure is non-increasing with respect to distance form the source. Similar to what has been mentioned in 2.2.1, Cuzick and Edwards (1990) proposed a test statistic that is the number of cases among the k nearest neighbors nearest the source of each case. On the contrary, Besag and Newell (1991) calculated the number of areas from the source required to accumulate k cases, as mentioned in 2.2.2.

In addition to the above methods, Diggle (1989,1991) proposed a method that first modeled a spatial process of a relatively more common disease that has no association with the source. Then, he compared this control process to the observed

spatial patterns of the disease of interest. The control process must be chosen so that there is no association between its distribution and the exposure to the source.

2.3 Detection of Clusters

In the early years, Choynowski (1959) proposed a quadrat-based test for the detection of cluster supplemented by a map on the brain tumors data in the Rzeszow province in Poland. There are 17 counties within the study region and each quadrat individually is tested whether the number of cases in it is significantly high based on a Poisson probability model. This method may not be able to detect clusters unless their boundaries coincide with the county borders.

Many other researchers have developed some cluster detection methods that relied on circles clusters, for example, GAM by Openshaw et al. (1988), CEPP by Turnbull (1990), Scan Statistic by Kulldorff and Nagarwalla (1995). These methods are designed by using multiple overlapping circles of variable radius as quadrats to increase the number of overlapped quadrats and thus overcome the problem of Choynowski's method. However, those methods may have lower power to detect non-circular clusters, Smith (2001) compared three methods (GAM/K, Kulldorff's spatial scan statistics, and Rushton and Lolonis'(1996) significance map) on a simulated sinuous cluster with a lower relative risk.

2.4 Discussion

In the discussion paper of Sankoh and Becher (2002), they suggested some appropriate disease cluster detection tests according to the purpose of the readers. Tackling group-level data in hand and clustering in space, methods recommended by Sankoh and Heiko Becher including Moran's I test, Besag and Newell's test, and Turnbull's test; besides that, Kulldorff and Nagarwalla's test, Openshaw's GAM, and Tango's test are also designed to handle group-level data.

As far as I know, most of the detection methods are designed to detect circular

clusters, therefore, they may have lower power to detect sinuous cluster. For the situation in Taiwan, each county is composed of numerous towns; most of these towns have neither regular shapes nor similar population densities. So, the test statistics mentioned above may not be useful if one is looking for insight into the sources of clustering.

Based on the geographical features in Taiwan and the aggregate data in hand, we have developed some procedures to detect the spatial pattern of the clusters, which are suspected to be in sinuous form along the rivers. Monte Carlo simulations are used to test the significance of the clusters.