Statistics for Biology and Health

For other titles published in this series, go to
http://www.springer.com/series/2848

Toshiro Tango

Statistical Methods for Disease Clustering

Springer

Toshiro Tango
Department of Technology Assesment & Biostatistics
National Institute of Public Health
3-6 Minami 2 chome
Wako, Saitama
351-0197 Japan
tango@niph.go.jp

Editors:

M. Gail
National Cancer Institute
Bethesda, MD 20892
USA

K. Krickeberg
Le Chatelet
F-63270 Manglieu
France

Jonathan M. Samet
Department of Preventive Medicine
Keck School of Medicine
University of Southern California
1441 Eastlake Ave. Room 4436, MC 9175
Los Angeles, CA 90089
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

W. Wong
Department of Statistics
Stanford University
Stanford, CA 94305-4065
USA

ISSN 1431-8776
ISBN 978-1-4419-1571-9 e-ISBN 978-1-4419-1572-6
DOI 10.1007/978-1-4419-1572-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010920016

Preface

This book is intended to provide a text on statistical methods for detecting clusters and/or clustering of health events that is of interest to final-year undergraduate- and graduate-level statistics, biostatistics, epidemiology, and geography students but will also be of relevance to public health practitioners, statisticians, biostatisticians, epidemiologists, medical geographers, human geographers, environmental scientists, and ecologists. Prerequisites are introductory biostatistics and epidemiology courses.

With increasing public health concerns about environmental risks, the need for sophisticated methods for analyzing spatial health events is immediate. Furthermore, the research area of statistical tests for disease clustering now attracts a wide audience due to the perceived need to implement wide-ranging monitoring systems to detect possible health-related bioterrorism activity. With this background and the development of the geographical information system (GIS), the analysis of disease clustering of health events has seen considerable development over the last decade. Therefore, several excellent books on spatial epidemiology and statistics have recently been published. However, it seems to me that there is no other book solely focusing on statistical methods for disease clustering. I hope that readers will find this book useful and interesting as an introduction to the subject.

Although the view of statistical methods of disease clustering embodied in this book is, of course, my own, it has been formed over many years through collaboration and contact with many statisticians. Especially, I must acknowledge the tremendous debt I owe to Martin Kulldorff, who has always provided me with invaluable insight and suggestions for improving my original ideas. I also thank Kunihiko Takahashi for preparing several figures and carefully reading the final text. My thanks also go to John Kimmel of Springer for inviting me to write this book and providing continual support and encouragement. Finally, I would like to thank Taeko Becque for checking my poor English.

Tokyo

Toshiro Tango
July 2009

Contents

Chapter 1

Introduction

In epidemiological studies, it is often of importance to evaluate whether a disease is randomly distributed or tends to occur as clusters over time and/or space after adjusting for known confounding factors, that may provide clues to the etiology of the disease. There has recently been great public concern about clustering of health events such as the occurrence of childhood leukemia, birth defects, and cancer. For example, since the early 1960s, it has been argued that childhood leukemia could be caused by either an infectious agent or an environmental toxin. Therefore, many researchers have examined space-time clustering of childhood leukemias in relation to the date and place of onset or diagnosis using various methods, including the Knox test. Furthermore, since the 1980s, there has been growing interest in the relation between the risk of a disease and proximity of residence to a prespecified putative source of hazard. It is well-known that the apparent excess of cases of childhood leukemia near a nuclear reprocessing plant such as that in the village near Seascale facility at Sellafield has been extensively investigated (for example, see Bithell *et al.*, 1994). More recently, there has been great public concern about the health effects of *dioxins*, organic compounds such as polychlorinated dibenzodioxins (PCDDs) and dibenzofurans (PCDFs), emitted from municipal solid waste incinerators (for example, see Elliott *et al.*, 1996).

In 1990, the Centers for Disease Control and Prevention (1990a, 1990b) issued the "Guidelines for investigating clusters of health events". In its appendix (1990b), a "summary of methods for statistically assessing clusters of health events" is provided as a resource for investigators who may become involved with the statistical aspect of *reported clusters* of health events.

In this book, I would like to introduce statistical methods for detecting disease clustering and/or localized clusters that are widely used and/or widely known in the literature and illustrate them with several real data sets. Almost all of the methods introduced here are used for *retrospective analysis*, except for the methods for syndromic surveillance in Chapter 9.

T. Tango, *Statistical Methods for Disease Clustering*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-1572-6_1,

1.1 Classification of Disease Clustering

Disease clustering is classified into one of three groups: *temporal clustering*, *spatial clustering*, or *space-time clustering*.

- *Temporal clustering* examines the question of whether cases tend to be located close to each other in time. One article illustrates this:

> An outbreak of acute nonbacterial gastroenteritis occurred among residents and staff in a nursing home in Baltimore, Maryland, in December 1980. A total of 101 residents and 69 staff members were surveyed by questionnaire. The attack rate (defined as acute onset of vomiting or two or more loose stools per 24 hours) was 46% of the group. Illness was brief and mild; no patients were hospitalized, and there were no deaths. Person-to-person transmission was documented by **temporal clustering** of cases (the demonstration of a higher rate of illness among residents exposed to an ill roommate one or two days earlier than among those not similarly exposed. ... The analysis of **temporal clustering** of cases was particularly useful in documenting person-to-person transmission in this outbreak and might be used for this purpose in other outbreaks caused by Norwalk or Norwalk-like viruses, as well as in outbreaks associated with other infectious organisms (Kaplan *et al.*, *American Journal of Epidemiology* 1982; **116**:940–948).

- *Spatial clustering* examines the question of whether cases tend to be located close to each other in space.
- *Space-time clustering* examines the question of whether cases that are close in space are also close in time. The following study illustrates this:

> The authors analyzed the natural history of multiple sclerosis (MS) before onset to identify the period of susceptibility and exogenous factors that might play a role in causing the disease. **Space-time cluster analysis** was performed among northern Sardinians, a genetically stable Italian population that showed an increasing risk of MS between 1965 and 1999. Residence changes from birth to clinical onset were recorded for all MS patients with clinical onset between 1965 and 1999 in the province of Sassari. ... **Clustering** was substantial in early childhood. **Clustering** was most marked in the most recent cases, among women, and among patients with early age at onset, a relapsing-remitting course, and in the eastern subarea... (Pugliatti *et al.*, *American Journal of Epidemiology* 2006; **164**:326–333).

To investigate whether clustering is real and significant, many different tests have been proposed for different purposes. Besag and Newell (1991) classified these tests into two families:

- *General tests* designed for investigating the question of whether clustering occurs over the study region.
- *Focused tests* designed for assessing the clustering around a pre-fixed point such as a nuclear installation. The following study illustrates this:

> Some recent epidemiologic studies suggest an association between lymphatic and hematopoietic cancers and residential exposure to high-frequency electromagnetic fields (100 kHz to 300 GHz) generated by radio and television transmitters. Vatican Radio is a very powerful station located in a northern suburb of Rome, Italy. In the 10-km area around the station, with 49,656 residents (in 1991), leukemia mortality among adults (aged $>$ 14 years; 40 cases) in 1987–1998 and childhood leukemia incidence (eight cases) in 1987–1999 were evaluated. The risk of childhood leukemia was higher than expected for distances up to 6 km from the radio station (standardized incidence rate = 2.2, 95% confidence interval: 1.0, 4.1), and **there was a significant decline in risk with increasing distance** both for male mortality ($p = 0.03$) and childhood leukemia ($p = 0.036$) (Michelozzi *et al.*, *American Journal of Epidemiology* 2002;**155**:1096–1103).

General tests were further classified by Kulldorff (1998) into two types:

- *Global clustering tests* designed for evaluating whether cases tend to come in groups or are located close to each other no matter when and where they occur. The following study illustrates this:

> A retrospective population-based case-control interview study has been conducted in three distinct areas in the north of England where local excesses of children with leukemia have been reported. A total of 109 cases of childhood (0–14 years at diagnosis) leukemia and non-Hodgkin's lymphoma who were born in one of the study areas and diagnosed there between 1974 and 1988 were included in the study. One control per case was matched on sex, date -of- birth, and health district of birth. **The objective was to compare residential histories of cases and controls and in particular to determine whether case children had lived in the same place at the same time more often than controls**. The residential distance between two children was taken to be the smallest geographical distance between homes they had "occupied" simultaneously for a period of at least six months between conception and diagnosis. **Case children were more likely than expected to have other cases as their nearest neighbors by residential distance** (observed = 69, expected = 54.5, p = 0.006) (Alexander *et al.*, *British Journal of Cancer* 1992; **65**:583–588).

- *Cluster detection tests* designed both for detecting localized clusters and evaluating their significance. The following article illustrates this:

> School immunization requirements are important in controlling vaccine-preventable diseases in the United States. Forty-eight states offer nonmedical exemptions to school immunization requirements. Children with exemptions are at increased risk of contracting and transmitting vaccine-preventable diseases. The clustering of nonmedical exemptions can affect a community's risk of vaccine-preventable diseases. The authors evaluated **spatial clustering** of nonmedical exemptions in Michigan and geographic overlap between exemption clusters and clusters of reported pertussis cases. **Kulldorff's scan statistic** identified 23 statistically **significant census tract clusters** for exemption rates and 6 **significant census tract clusters** for reported pertussis cases between 1993 and 2004 (Omer *et al.*, *American Journal of Epidemiology* 2008;**168**:1389–1396).

1.2 Data Used for Disease Clustering

It is probably no exaggeration to say that the cluster investigation started to explore space-time clustering of childhood leukemia in the mid 1960s, when Knox's test (1964a, 1964b) and Mantel's test (1967) were applied. To test for disease clustering, we generally need to collect data retrospectively on the *time of occurrence* and/or the *location* of each case for a defined geographic region during a specified study period. *Date of onset*, *date of birth*, and *date of death* have been used as the time of occurrence for clustering of childhood leukemia. The rationale for using date of onset invokes a short time interval between the inductive event and onset of disease. A search for clustering by date of birth invokes the hypothesis that childhood leukemias are determined pre- and perinatally, when the human organism is particularly susceptible to the effects of a carcinogen. Needless to say, very great care is required for clustering by date of death because of the long and variable interval between the onset and death. *Address at onset*, *address at birth*, and *address at death* can be used as the *location* depending on the study objectives. Examples of data used to explore space-time clustering of childhood leukemia are shown below:

Onset cluster: Klauber and Mustacchi (1970) investigated space-time clustering for *time of diagnosis* and address of 149 leukemia cases in children under the age of 15 years diagnosed in San Francisco during the 20-year period 1946 to 1965.
Birth cluster: Klauber (1968) conducted a study of clustering of childhood leukemia by *hospital of birth* where 234 children under age 5 who died of leukemia and who were born during 1958–1960 in California were identified.
Death cluster: Glass and Mantel (1969) analyzed dates and residence data for 298 Los Angeles childhood leukemia deaths during the period 1960–1964.

Data types used for cluster investigation, on the other hand, are generally classified into two types:

- *Individual geographical location data* include data on coordinates of birth, residence, workplace, and death, which are usually residential addresses such as street address, zip code, and post code unit.
- *Regional count data* include the number of cases and population at risk in some small areas usually defined for administrative purposes, such as census tracts, counties, municipalities, and electoral wards.

Due to the restriction of clinical confidentiality, it is often impossible to obtain individual data, and so we have to resort to an analysis of the count within an administrative region.

1.3 Organization of the Book

Chapters 2 and 3 provide readers with a brief introduction to basic concepts of *clustering* and *clusters* and the basic idea of how to detect clustering and disease mapping, which are useful tools for visualizing the *clustering* and *clusters* before going into the details of statistical tests. The remaining chapters are arranged according to the particular type of disease clustering. The book is organized as follows:

- Chapter 2 introduces to the basic concepts of *clustering* and *clusters* and the basic idea of how to detect clustering and/or clusters.
- Chapter 3 introduces to basic concepts of disease mapping and a range of mapping methods that are useful tools for visualizing regional variations of disease risk or regional clustering and/or clusters.
- Chapter 4 presents tests for detecting temporal clustering.
- Chapter 5 discusses general tests for detecting spatial clustering based on regional count data.
- Chapter 6 gives general tests for detecting spatial clustering based on case-control location data.
- Chapter 7 gives tests for space-time clustering or space-time interaction.
- Chapter 8 discusses focused tests for spatial clustering.
- Chapter 9 discusses the space-time scan statistic with special emphasis on the application to a syndromic surveillance.

1.4 Organization of the Chapters

Each of Chapters 4 through 9 has the following basic structure:

- Each chapter begins with illustrative real examples that are later analyzed by some of the selected methods.

- The *Data* section describes the data necessary for cluster investigation and the notation used throughout the chapter.
- The *Null Hypothesis vs. Alternative Hypothesis* section describes the null hypothesis and alternative hypothesis both nonstatistically and statistically.
- The *Historical Overview of Methods* section gives an overview of the *major* methods proposed so far and is not a review of all the methods. So, readers should note that there are several other proposed methods not mentioned here. Furthermore, readers who are not interested in the history can skip this section.
- The *Selected Methods* section describes the details of selected methods, most of which are widely known and/or widely used.
- The *Illustration with Real Data* section illustrates the selected methods with real data sets.
- The *Power Comparison* section compares the powers of the selected methods.
- The *Discussion* section discusses the appropriateness and the relative merits of the methods. This section is generally provided when the *Power Comparison* section could not be set.

1.5 Statistical Software

In each of Chapters 4 through 9, selected methods for detecting disease clustering are illustrated with real data in the *Illustration with Real Data* section using some of the software packages listed below, which are available free of charge. In particular, in the appendix, I provide readers with R functions for some selected methods for ease of applying the methods.

1.5.1 R

URL: http://www.r-project.org/

R is a language and environment for statistical computing and graphics. It is a GNU project that is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T) by John Chambers and colleagues. R can be considered a different implementation of S. There are some important differences, but much code written for S runs unaltered under R (see the *Introduction to R*).

1.5.2 SaTScan

URL: http://www.satscan.org/

SaTScan is free software that analyzes spatial, temporal, and space-time data using the spatial, temporal, or space-time scan statistics. It is designed for any of the following interrelated purposes:

- To perform geographical surveillance of disease, detect spatial or space-time disease clusters, and see if they are statistically significant.
- To test whether a disease is randomly distributed over space, time, or space and time.
- To evaluate the statistical significance of disease cluster alarms.
- To perform repeated time-periodic disease surveillance for early detection of disease outbreaks.

1.5.3 FleXScan

URL: http://www.niph.go.jp/soshiki/gijutsu/download/flexscan/index_e.html

FlexScan is free software developed to analyze spatial count data using the flexible spatial scan statistic and circular spatial scan statistic. The current version includes a spatial scan statistic with a restricted likelihood ratio. These scan statistics are introduced in Chapter 5. FLeXScan is similar to SaTScan, but the current version of FleXScan is still restricted to spatial analyses.

1.5.4 Splancs

URL: http://www.maths.lancs.ac.uk/rowlings/Splancs/

Splancs is a software package for spatial and space-time point pattern analysis (Rowlingson and Diggle, 1993) that can be installed in R. See also Bivand and Gebhardt (2000).

Chapter 2
Clustering and Clusters

In this chapter, we shall show several spatial point patterns in a hypothetical area of a square 10 km on each side to understand what "clustering" or "a cluster" means and to get a basic idea of how to approach detection of *clustering* or *clusters*.

2.1 Spatial Pattern

Figure 2.1 shows a *random pattern* of spatial locations of 100 cases within a square where the spatial location of each case is completely independent of the spatial location of every other case. This random pattern is also called *complete spatial*

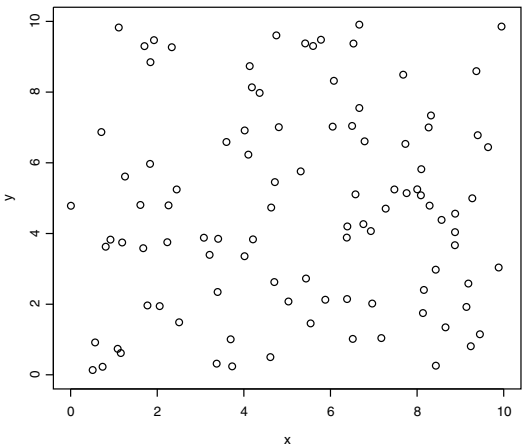

Fig. 2.1 A *random* pattern of spatial locations of 100 cases.

T. Tango, *Statistical Methods for Disease Clustering*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-1572-6_2,

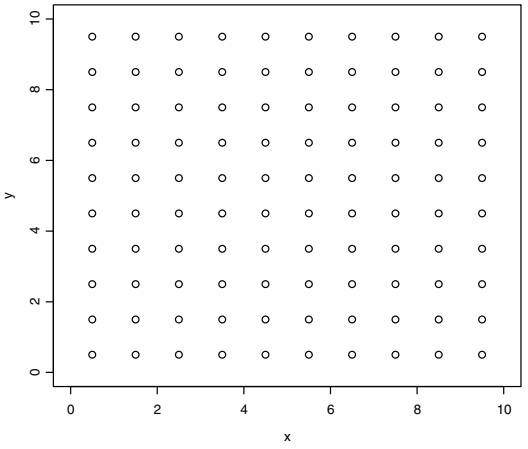

Fig. 2.2 A *completely regular* pattern of spatial locations of 100 cases.

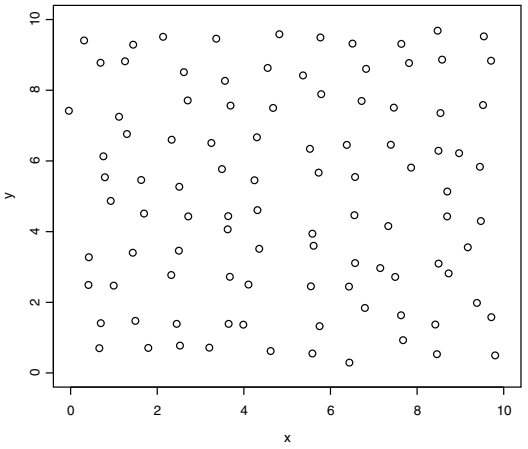

Fig. 2.3 A *regular* pattern of spatial locations of 100 cases.

randomness in the literature and is widely used as the null hypothesis of no clustering for statistical hypothesis testing. Although we can see that some points are aggregated here and there, these apparent aggregations are made up of a complete random mechanism. On the other hand, Figure 2.2 shows a completely *regular* pattern of spatial locations of 100 cases within a square where every case is the same distance from its nearest neighbor. Then, what is Figure 2.3? Is it a random pattern?

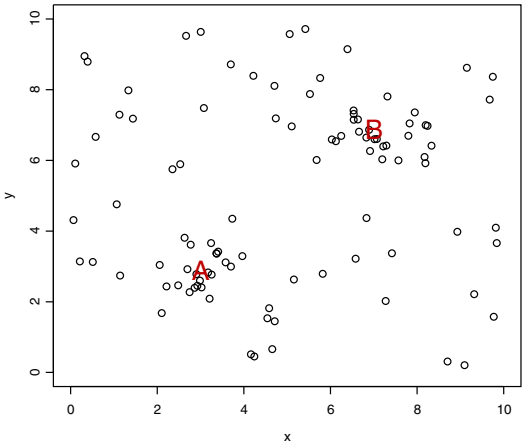

Fig. 2.4 An *aggregated* pattern (Pattern I) of spatial locations of 100 cases.

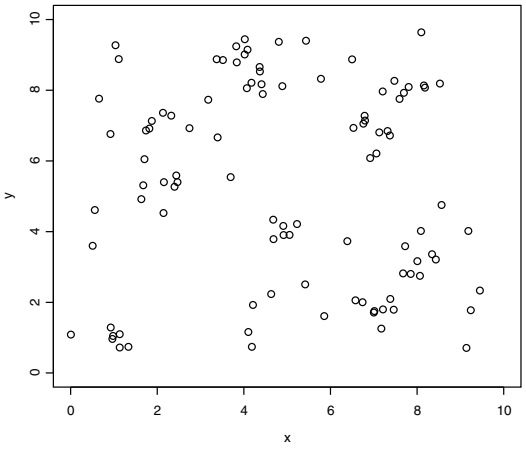

Fig. 2.5 Another *aggregated* pattern (Pattern II) of spatial locations of 100 cases.

No, it is still a *regular* pattern in the sense that cases are more spaced out than in a random pattern as in Figure 2.1. Figures 2.4 and 2.5 show two kinds of *aggregated* patterns of spatial locations, Pattern I and Pattern II, where cases are more aggregated than in a random pattern. In Figure 2.4, you may observe that cases are aggregated to a certain extent around two areas, A and B. In this situation, we call a set of aggregated cases *a cluster of cases* or *clustered cases*, and we may say

that *there are two apparent clusters of cases* in this area. Some readers might be interested in the following questions:

- Is there significant clustering of cases in this area?
- How do you estimate the location of localized clusters and test their statistical significance?

Figure 2.5, on the other hand, shows that there seem to be *many small clusters of cases throughout the area*. If the disease under study is infectious, we would expect cases to be found close to each other no matter where they occur, as in Figure 2.5. Therefore, in this situation, we are not interested in localized clusters but in the question

- Is there significant clustering of cases in this area?

Before considering these questions, I would like to briefly introduce statistical models of the *spatial point process* that generates spatial point patterns.

2.2 Spatial Point Process

2.2.1 Homogeneous Poisson Process

A spatial point process describes a probabilistic model where each random variable represents the location of an event in space. In particular, a stationary and isotropic *homogeneous Poisson process* is very important and is defined by the following criteria (for example, see Diggle, 2003; Waller and Gotway, 2004, Section 5.2):

1. *Stationarity* requires that a process be invariant to translation within space.
2. *Isotropy* requires that a process be invariant to rotation about the origin.
3. The number of events occurring within an area A is a random variable following a Poisson distribution with mean $\lambda \mid A \mid$, where λ is called *intensity* and $\mid A \mid$ denotes the area of A.
4. Given the total number of events n occurring within an area A, the n events represent an independent random sample of n *locations*, each event uniformly distributed over the area.

Both stationarity and isotropy mean that the relationship between two events depends only on the distance between them. The intensity λ of the process introduced in criterion 3 means the *average number of points per unit area*; i.e., it is *estimated* by

$$\hat{\lambda} = \frac{\text{the number of events in A}}{\mid A \mid} \tag{2.1}$$

where the intensity is assumed to be constant at all locations and thus the process is called *homogeneous*. It should be noted that the "hat" over the parameter, λ here,

defines an *estimate* throughout the book. Criteria 3 and 4 indicate that the numbers of events in disjoint regions are statistically independent, which is an important property for the analysis of regional count data. Finally, it should be noted that the definition above describes the statistical model for *complete spatial randomness*. The intensity is also called the *first-order* measure of a spatial point process since it describes the *mean* of the process. More useful for evaluating clustering in space is one of the *second-order* measures related to the *variance* of the process, called Ripley's K-function (Ripley, 1976, 1977), which is defined as

$$K(s) = \frac{E[\text{Number of further events within distance } s \text{ of an arbitrary event}]}{\lambda} \tag{2.2}$$

Under the null hypothesis of no clustering or a stationary and isotropic homogeneous Poisson process, the expected number of events within distance s of an arbitrary event is $\lambda\pi s^2$. Therefore, if there is clustering, then we expect an excess of events at short distance; i.e., $K(s) > \pi s^2$ for small s. If the boundary of the area A is sufficiently far away from all the points, then the K-function is usually calculated as

$$\hat{K}(s) = \frac{|A|}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} I(d_{ij} \leq s)$$

where d_{ij} denotes the Euclidean distance between events i and j and $I(.)$ is the indicator function. However, in this book, we shall use the definition

$$\hat{K}(s) = \frac{|A|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} I(d_{ij} \leq s) \tag{2.3}$$

which gives an *unbiased* estimator of the K-function (Diggle and Chetwynd, 1991). In the study of spatial point patterns generated by a homogeneous Poisson process, investigators usually have to observe the spatial pattern only in a limited area, but their major purpose is the estimation of the nature and second-order properties of the spatial patterns such as the locations of trees in a forest and cell nuclei in a microscopic tissue section rather than the clustering of events for small s. Therefore, the formula (2.3) is not appropriate for that purpose because it would not include events occurring outside the area A for s larger than the distance of an event to the nearest boundary or edge. As a method to cope with this problem, the *edge-corrected* estimator

$$\hat{K}(s) = \frac{|A|}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} w_{ij} I(d_{ij} \leq s) \tag{2.4}$$

is usually used, where w_{ij} is defined as the reciprocal of the proportion of the circumference of the circle centered at an event i with radius d_{ij} that lies within A. In other words, w_{ij} is the reciprocal of the conditional probability that an event j falls within the study region given that its distance to the event i is d_{ij}.

In the study of disease clustering, on the other hand, we are generally not interested in data outside the study area. Therefore, the K-function estimated without the edge-correction term is enough. In this book, we introduce several tests for disease clustering based on the K-function with the edge-correction term. However, when applying these methods, we have only to draw some *larger* boundary of the study region A that makes the difference between the K-function with and without the edge-correction term negligible.

2.2.2 Inhomogeneous Poisson Process

However, it is unnatural to assume a homogeneous Poisson process for examining disease clustering because the intensity strongly depends on the density of the population at risk. Therefore, we have to introduce an *inhomogeneous Poisson process* with a spatially varying intensity $\lambda(\boldsymbol{z})$ (at location $\boldsymbol{z}$), defined as:

1. The number of events occurring within a region $D \subset A$ is a random variable following a Poisson distribution with mean $\int_D \lambda(\boldsymbol{x})d\boldsymbol{z}$.
2. Given the total number of events n occurring within the study area A, the n events represent an independent random sample of n *locations* with the probability of sampling a particular point $\boldsymbol{z}$ proportional to $\lambda(\boldsymbol{z})$.

where the intensity is at least a function of the population density. However, generally the population density is not known or is difficult to compute on a local scale. Therefore, it is not practical to apply an inhomogeneous Poisson process directly to individual point data for examining disease clustering and instead we have to devise some method of escaping this difficulty.

2.3 Back to the Questions

Let us assume that the spatial distribution of the population at risk in our hypothetical area is as shown in Figure 2.6, indicating that there are two large clusters of populations at the locations similar to those of the clustered cases shown in Figure 2.4. Therefore, two apparent clusters of cases in *Pattern I* might just be due to the fact that these two areas A and B are densely populated areas. So, the problem here, based on these two types of data, the case locations and the population (or control) locations, is what kind of approach we can take to testing the null hypothesis H_0 of no clustering against the alternative hypothesis H_1:

$$H_0 : \text{there is no clustering of cases in this area}$$
$$H_1 : \text{there is clustering of cases in this area}$$

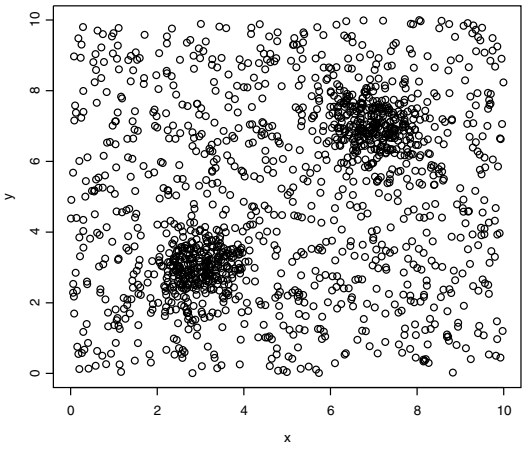

Fig. 2.6 Spatial locations of the population at risk (1500 residents).

2.4 Approaches Using Regional Count Data

If this area is partitioned into several *administrative regions* such as census tracts and block groups, an easy-to-understand and naive approach is to *count* the number of cases for each region and compare a set of *counts observed* with the corresponding set of *counts expected under the null hypothesis*. Then, let us suppose that this hypothetical area is partitioned into 25 regions or squares with each side 2 km, shown in Figure 2.7, and examine the two patterns of locations of cases above.

Example 1. Pattern I

The points A and B in Figure 2.4 are located in the regions No.7 and No. 19, respectively. Then, the observed number of cases $O_i, i = 1,...,25$ and the population at risk $\xi_i, i = 1,...,25$ per region are shown in Figure 2.8 and Figure 2.9, respectively. You can see that two regions, No.7 and No.19, have a large number of cases and large population at risk compared with other regions. So, to see if these two regions have higher rates than other regions, let us compute the observed proportion or rate per region (Figure 2.10). However, we cannot observe any high rates because the average rate is 0.064. An alternative way to look at the peculiarity of each region is to use the ratio of the *observed* (O) to the *expected* (E) number of cases, called the O/E ratio or *standardized mortality (incidence) ratio* (SMR) . The SMR in region i is defined as

$$SMR_i = \frac{O_i}{E_i} \tag{2.5}$$

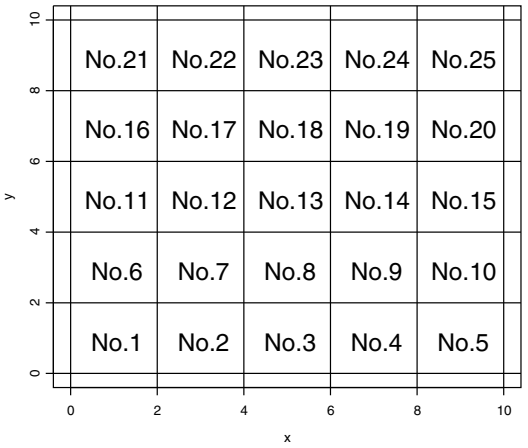

Fig. 2.7 Locations of 25 squares with each side 2 km.

where the expected number of cases E_i is calculated under the null hypothesis of no clustering. Although SMR usually refers to mortality rather than incidence, the term SMR is widely used for both mortality and morbidity, including incidence. Therefore, we shall here use the term SMR irrespective of mortality or incidence. To calculate the number of cases expected in each of 25 regions, we can apply the property of the inhomogeneous Poisson process that tells us that the number of events observed in disjoint regions still follows an independent Poisson distribution where the expected number of events in region D is proportional to $\int_D \lambda(\boldsymbol{x})d\boldsymbol{x}$. Namely, if the null hypothesis is true, the expected number of cases is proportional to the population at risk. Given the total number of cases $n = O_1 + \cdots + O_{25}$, the expected number of cases is obtained by

$$\begin{aligned} E_i &= n \times \text{the ratio of the } i\text{th region's population to the total population} \\ &= n \times \frac{\xi_i}{\sum_{j=1}^{25} \xi_j} \end{aligned} \tag{2.6}$$

In this case, we have the following equality among the E_i's:

$$\sum_{i=1}^{25} E_i = n \tag{2.7}$$

Figure 2.11 shows SMR values for each region. SMR values greater than 1.0 indicate more cases observed than expected. The SMRs in regions No.7 and No.19 are 1.2 and 1.1, respectively, larger than 1.0 but not so large. Let us go back to the

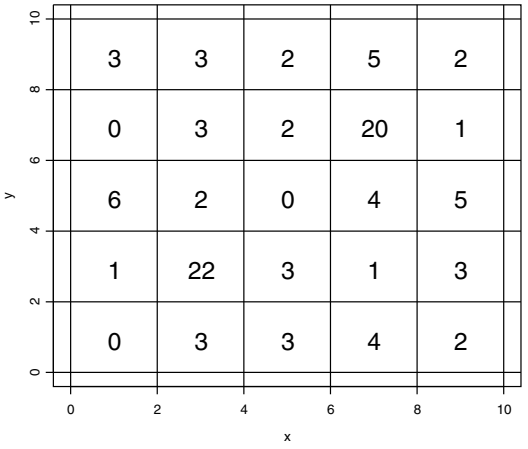

Fig. 2.8 The observed number of cases in each of 25 squares.

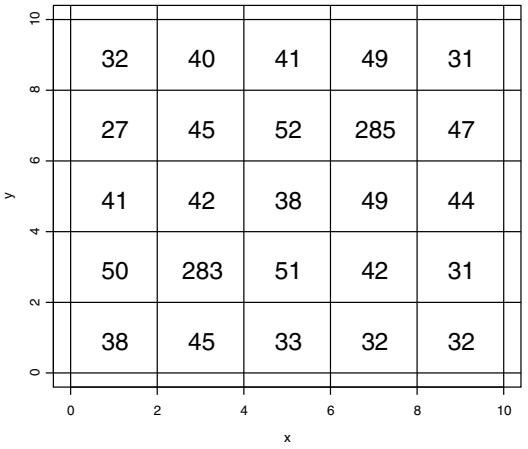

Fig. 2.9 Population at risk in each of 25 squares.

problem of testing the null hypothesis H_0 of no clustering. If we observe a number of independent cases in each region in the study area, these data can usually be analyzed by comparing the frequency distribution of counts with a Poisson distribution. If the observed number of cases O_i in region i follows independent Poisson random variables with the expected value E_i and variance E_i, then the standardized residual is given by the Z-value

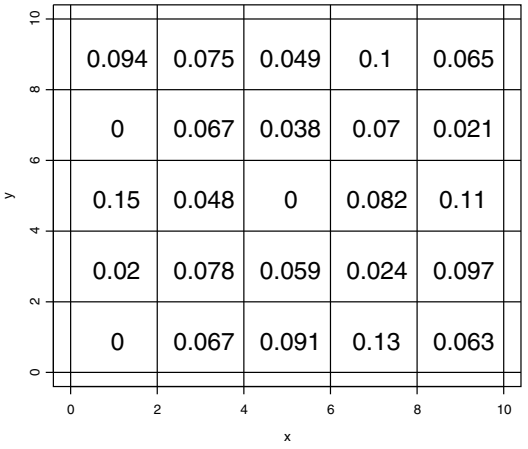

Fig. 2.10 Observed proportions of cases in each of 25 squares.

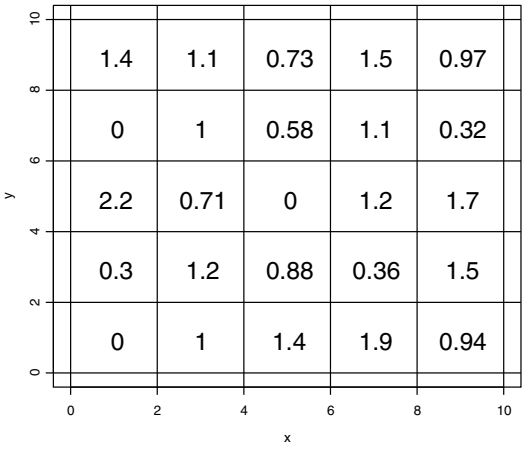

Fig. 2.11 Standardized mortality ratio in each of 25 squares.

$$Z_i = \frac{O_i - E_i}{\sqrt{E_i}} \tag{2.8}$$

which has approximately a standard Normal distribution $N(0,1)$ with mean zero and variance 1. Z-values larger than 2.0 or less than -2.0, say, suggest outlying high or low values. Figure 2.12 shows the individual Z-value, indicating that neither region No.7 with $Z = 0.72$ nor No.19 with $Z = 0.23$ are outlying. It should be noted