



A brief evaluation of statistical methods for  
detecting disease clusters in time and/or space

**MATH 4779/5779 Fall 2004**

Department of Mathematics  
University of Colorado-Denver

Karen Kafadar, adviser

Andy K. Kim, research assistant

Sami A. Al-Rumaizan

Corey J. Ayala

Andrew J. Been

Jingjing Chen

Jason Cook

Gustavo G. Duarte

Catherine S. Durso

Jason R. Gonzales

Co N. Huynh

Tu V. Huynh

Ilya V. Lashuk

Diane M. Wagner

Salvador S. Sanabria

November 18, 2004

<sup>1</sup>Sponsored by Colorado Department of Public Health and the Environment  
(CDPHE)

# Contents

# Chapter 1

## Introduction

### 1.1 Statement of problem

The timely detection of potential outbreaks of serious communicable diseases is one of the critical functions of public health departments, both at the state and federal level. Because many communicable diseases are required to be reported to local health departments, Colorado Department of Public Health and the Environment (CDPHE) receives notification of cases from epidemiologists, hospitals, physicians' offices, and various other health facility monitors. CDPHE maintains a database which includes, for each case, the diagnosed disease, temporal information (e.g., date of occurrence, date of report), and spatial information (e.g., population centroid of the census tract where the diseased person live, suspected location where the disease may have been acquired). When an outbreak occurs due to a communicable disease, the first priority of public health personnel is to reduce mortality and morbidity. Thus, prompt and immediate detection of cases of disease clusters (e.g., many cases in a short period of time, and/or many cases in a rather small geographic area), is an urgent priority for CDPHE.

Public health departments have been conducting surveillance activities using various methods for decades. These activities have become particularly important in Colorado during the past several years, for several reasons. Since September 11, 2001, the potential for bioterrorism has increased, requiring increased vigilance among health departments to check for unusually high incidence of unexpected diseases (disease surveillance) or symptoms (syndromic surveillance). The identification of anthrax on the East Coast shortly after 9/11/01 emphasized the awareness of the need to monitor public health databases for suspected bioterrorism activities. Another event that has hit Colorado particularly hard had been the spread of West Nile Virus;

last year, Colorado led the 50 states in the most number of reported cases, despite the fact that it ranks 22nd in total population. The outbreak of SARS in Asia last year underscored the importance of extremely rapid identification and containment of a communicable disease outbreak, particularly in a state like Colorado that depends heavily on tourism year-round.

How can CDHPE's database be used to detect potential outbreaks of disease? Statistical methods for cluster detection can aid in this identification. All cluster detection methods run the risk of two types of error: the error in claiming an outbreak when the incidence is in fact below "outbreak" levels (here, called Type I error), and the error in failing to detect an outbreak when one exists (here, called the Type II error). The first cost arises from the mobilized resources in increased personnel, material, and public fear. The second cost arises from the potentially large numbers of cases that might have been prevented if the outbreak had been detected.

Statistical methods are valuable tools in detecting disease outbreaks, because error rates from these two types of error can be quantified mathematically. The proper specification of these error rates depends upon the assumptions of the methods, usually the assumptions governing what constitutes "below outbreak levels."

1. *Temporal Clustering (Chen, Hyunh)*

The lack of disease clusters in time assumes that cases of disease occur uniformly in time; i.e., the observed number of cases of a particular disease is about the same in one interval as in previous or succeeding intervals. Because of the seasonality of many disease (e.g., influenza), it is important that the time intervals under consideration are not too wide. The Centers for Disease Control and Prevention has addressed this issue by comparing the number of diseases in the current four-week period to the numbers reported in previous four-week periods, during the previous five years, from the same four-week period as well as exactly one four-week period on either side of it (i.e., comparing the current count to 15 previous counts). To the extent that this count is "significantly higher" than the average of the 15 previous counts, time clustering is indicated. (See Section 7 in Chapter 5 for more details on this method.) Other statistical methods for detecting temporal clusters assume basically zero counts; the existence of a positive count in one or more intervals may lead to "significance," depending upon the statistical method.

2. *Spatial Clustering (Been, Gonzales, Hyunh, Wagner, Durso)*

The lack of disease clusters in space assumes that cases of disease occur

randomly, or uniformly, over the region of interest; i.e., the observed number of cases of a particular disease is about the same in one region as in other regions, whether they be close by or far away. Because the number of cases in a region obviously depends upon the number of people who live there, these spatial counts must be reported either as *rates* (e.g., number of cases per 1000 persons), or in regions that are designed to contain roughly the same numbers of persons (e.g, census tracts within county borders). Many statistical methods have been proposed for detecting spatial clusters. Some methods (e.g., Knox’s method, Kulldorff’s Spatial Scan Statistic) consider clusters in both space and time (i.e., an unusually high number of cases in a small region and in a short period of time).

### 3. *Hypothesis test and Errors (Kim)*

All detection methods depend upon the criteria which describe the "no outbreak" situation (i.e., the "null hypothesis" and no urgent action need be taken). From this hypothesis, the error rates (i.e., declaring an outbreak when none exists, or failing to declare an outbreak when it exists), can be quantified. The criteria for the test depend upon appropriate specification of one of these two types of error. For example, CDPHE may want to specify no more than a 5% chance of dispatching personnel and resources for a non-urgent condition. Alternatively, the CDPHE may prefer to specify no more than a 1% chance that an outbreak will fail to be detected. Given the specification, the goal is to choose a method that (a) minimize the other error rate (e.g., if the Type I error is specified at 5%, then the method maintains as small a Type II error rate as possible; or, if the Type II error is specified at 1%, then the method maintains as small a Type I error rate as possible); and (b) depends on assumptions that are reasonable for CDPHE’s situations. Practical reasons dictate the infeasibility of the second approach (due to the number of possible forms that an outbreak can take, in space and/or time, it would be nearly impossible to assure no more than a 1% chance that an outbreak of *any* form would fail to be detected). There is always a trade-off between the two types of error: a very small Type I error (chance of declaring an outbreak when none has arisen) leads to a large Type II error (chance of failing to notice an outbreak when one occurs), and vice versa, so CDPHE may well wish to apply statistical tests with a Type I error rate that is higher than the conventional 5%. On the other hand, many tests for clustering also leads to more chance to commit Type I errors; repeating 5% tests 20

times ensures, by construction, the likelihood of declaring at least one outbreak when none exists (5% chance per test, times 20 tests, equals 100% expected chance of declaring an false outbreak).

## **1.2 Executive Summary**

This report describes various statistical methods that have been proposed in the literature and critically assesses them for their consistency with reasonable assumptions, for their ease in computing the methods, and for their ability to maintain low error rates of the two types. Following this introductory chapter, Chapter 2 gives a general description of one particular disease of concern to Colorado (West Nile Virus). Chapter 3 provides some background on classical hypothesis testing. Chapter 4 describes a de-identified data set containing lists of cases, locations, and times of occurrence. Chapters 5, 6, and 7 describe statistical tests for detecting clusters in time, space, and space and time, respectively. Throughout these chapters, selected methods are illustrated using the de-identified data set. Chapter 8 provides some recommendations on displaying the data. Chapter 9 summarizes this report with recommendations on methods for temporal and/or spatial clustering, along with illustrations using the CDPHE data set.

# Chapter 2

## Epidemiology

### 2.1 History

West Nile virus was first isolated from a febrile adult woman in the West Nile District of Uganda in 1937 [?, ?, ?]. The ecology was characterized in Egypt in the 1950s [?]. The virus became recognized as a cause of severe human meningitis or encephalitis (inflammation of the spinal cord and brain) in elderly patients during an outbreak in Israel in 1957 [?]. Equine disease was noted first in Egypt and France in the early 1960s [?]. In the mid 1990s outbreaks were reported from Romania, Russia and Israel [?].

WNV first appeared in North America in 1999, with encephalitis reported in humans and horses [?]. Although the virus spread westward during the next two years, only modest disease activity was seen until 2002, when the number of cases increased dramatically, and by the end of 2003 the epizootic had spread to all but two of the lower 48 states [?]. The number of cases of WNV has also continued to rise; by mid-February 2004, 9175 human cases and 230 deaths were reported as a result of the 2003 outbreak [?].

### 2.2 Geographic distribution

West Nile virus has been detected in Africa, Europe, the Middle East, west and central Asia, Oceania (subtype Kunjin), and most recently, North America[?, ?].

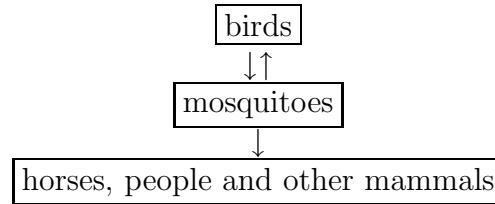


Figure 2.1: WNV is maintained in an enzootic mosquito-bird-mosquito cycle. Humans and other mammals serve as dead-end hosts and do not sufficiently amplify virus for mosquito transmission.

Outbreaks of WNV encephalitis in humans have occurred in Algeria in 1994, Romania in 1996-1997, the Czech Republic in 1997, the Democratic Republic of the Congo in 1998, Russia in 1999, the United States in 1999-2003, and Israel in 2000. Epizootics of disease in horses occurred in Morocco in 1996, Italy in 1998, the United States in 1999-2001, and France in 2000, and in birds in Israel in 1997-2001 and in the United States in 1999-2002 [?, ?, ?].

## 2.3 Virology

WN virus is taxonomically placed within the family Flaviviridae, genus Flavivirus. Within the genus Flavivirus, WN virus has been serologically classified within the JE virus antigenic complex, which includes the human pathogens JE, Murray Valley encephalitis, SLE, and Kunjin viruses [?].

## 2.4 Transmission cycle

WNV is maintained in an enzootic bird-mosquito-bird cycle; see figure ???. The virus is amplified during periods of adult mosquito blood-feeding by continuous transmission between mosquito vectors and bird reservoir hosts [?].

In Africa, southern Europe, and western Asia, WNV has been isolated from mosquitoes of more than 40 species, primarily those in the genus *Culex*; in the United States, WNV has been isolated from mosquitoes belonging to 43 species since 1999. In the United States the majority

of WNV isolates have been from *Culex* species, particularly *Cx. pipiens*, *Cx. restuans*, and *Cx. salinarius*; however, isolates have also been recovered from species in other genera, including *Aedes*, *Anopheles*, *Coquillettidia*, and *Ochlerotatus*. *Culex* species are the most important maintenance vectors within the avian cycle, with other species serving as bridge vectors from birds to humans and horses in mid to late summer. While mosquitoes belonging to many species are competent vectors of the virus in the laboratory, their vector competence in humans is still unknown. In the northeastern and southern US transmission among birds is likely mediated by *Cx. pipiens*; however, the most important bridge vectors to humans are unidentified. The large outbreak of WNV in the western US, especially Colorado, in 2003 was likely due to *Cx. tarsalis*, an indiscriminant feeder on both birds and mammals. In regions where the same vector contributes to both the mammalian and the avian cycles, disease activity will continue to be more pronounced. WNV has also been isolated from both hard and soft ticks; however, the ability of ticks to successfully and significantly transmit the virus in nature is unclear [?].

Passerine birds, including crows, house sparrows, and blue jays, serve as the primary amplifying hosts of the virus, and develop a high-level viremia that lasts for several days. A study of WNV transmission in 25 species of birds found cloacal shedding of virus in 17 of 24 species and oral shedding in 12 of 14 species. In addition, contact transmission was identified in four species, and oral transmission in five species [?].

Infectious mosquitoes carry virus particles in their salivary glands and infect susceptible bird species during blood-meal feeding. Competent bird reservoirs will sustain an infectious viremia (virus circulating in the bloodstream) for 1 to 4 days after exposure, after which these hosts develop life-long immunity. A sufficient number of vectors must feed on an infectious host to ensure that some survive long enough to feed again on a susceptible reservoir host [?].

Most human infections of WNV are the result of transmission of the virus by infected mosquitoes; however, several novel modes of transmission are now recognized in the United States (in utero, through breast milk, via blood transfusion or organ transplantation, or through occupational exposure). In contrast to some bird species, in which high-level viremia is seen, humans and horses develop only low-level and transient viremia and are unlikely to contribute to virus amplification via mammal-mosquito-anywho transmission cycle [?].

## 2.5 Clinical features

Most infections with WNV are clinically inapparent and go undetected. A serosurvey in 1999 in New York showed that only approximately 20% of infected persons developed fever caused by WNV, and of these, only about half visited a physician for their illness. Approximately 1 in 150 patients progress to severe neurologic illnesses (encephalitis, meningitis, acute flaccid paralysis), and while prevalence rates are fairly uniform across age-groups, rates of neurological disease increase substantially with age, as does the clinical:subclinical infection ratio. Case-fatality rates among hospitalized patients have ranged from 4% (Romania, 1996) to 12% in the New York 1999 outbreak and as high as 14% in an outbreak in Israel in 2000. Advanced age is the main risk factor for death, with individuals more than 70 years of age at particularly high risk. Among such individuals, the case-fatality rate ranges from 15% to 29%. The reason for increased mortality in the elderly is not yet known but may be related to a decreased capacity of these individuals to develop a protective immune response to help control infection [?].

Currently, the only treatments for WNV infection are supportive [?].

## 2.6 WNV in Colorado

According to [?] and [?], there was no evidence of WNV activity in Colorado until 2001. In 2001 several reports about human and veterinary cases were received, but none of the samples submitted for testing was found to be positive for WNV (see [?]). In 2002 there were reported and confirmed human, veterinary and wild bird cases (see [?, ?]). In 2003 there were reported 2947 human cases (according to [?]), the highest value among all the states for 2003. This year there were reported only 276 human cases (data from [?]). In both 2003 and 2004 there were reported and confirmed veterinary and wild bird cases ([?, ?]).

# Chapter 3

## Statistical Hypothesis Test

There are two general methods available for making inferences about the population parameters, or in our situation, the existence of disease clusters. We can estimate their values using confidence intervals, or we can make decisions about them. However, when the inference leads to a decision about certain actions or certain phenomenon, rather than specific values of model parameters that may be related to the conclusion, we prefer the decision making process, or statistical hypothesis-testing. A statistical hypothesis test, or more briefly, hypothesis test, states two alternatives of possible outcomes, and decides the likely outcome based on the observed data and on specified risks of making the wrong choice.

### 3.1 Elements of Hypothesis Testing

Every hypothesis test consists of five fundamental elements.

- (a)  $H_0$  (Null hypothesis)
- (b)  $H_a$  (Alternative hypothesis)
- (c) test statistic
- (d) rejection region (or “critical region”)
- (e) decision algorithm (based on test statistic)

### 3.1.1 Null and Alternative Hypothesis

Before collecting information or data about the phenomenon of interest (e.g., existence of clusters), the researcher formulates hypotheses that could describe the state of nature. One hypothesis usually describes the current state of nature, in which nothing very interesting or alarming is happening. This situation is usually captured in the “null hypothesis.” In our situation, the null hypothesis might be, “no evidence of clustering.” Conversely, the “alternative hypothesis” captures the unusual or the unexpected; e.g., “evidence of clustering.”

The null hypothesis is denoted  $H_0$ ; it describes the current state of affairs, and is usually the hypothesis that the researcher aims to dispel.

The alternative hypothesis is denoted  $H_a$ ; it describes an unexpected departure from the current state, and evidence supporting it is usually considered important or critical.

Thus, the null and alternative hypotheses describe the population, but from different and mutually exclusive perspectives (i.e., they cannot be true simultaneously). This is important, because it implies the conditions governing the population can be described by either  $H_a$  or  $H_0$ , but not both. (In some contexts, a third “middle” option is possible, which describes the situation where the evidence is insufficient to make a determination on way or the other. In such “sequential testing” paradigms, data continue to be collected until the evidence points convincingly to either  $H_0$  or  $H_a$ .)

A classical example of hypothesis testing is a court case. Here, the prosecutor may be thought of as the researcher who seek to find the evidence to prove the defendant guilty. Our judicial system is set up on the premise that the defendant is innocent ( $H_0$ ) until proven guilty “beyond a shadow of a doubt” ( $H_a$ ). Thus,

$H_0$  : the defendant is not guilty

$H_a$  : the defendant is guilty

Until proven guilty, the defendant remains not guilty ( $H_0$ ).

In our context,

$H_0$  : no evidence of disease clusters in the region; e.g., cases are distributed randomly throughout the region

$H_a$  : evidence of disease cluster

### 3.1.2 Type I and Type II Errors

The goal of any hypothesis test is to make a decision. In particular, we will decide whether to reject the null hypothesis  $H_0$  in favor of the alternative hypothesis  $H_a$ . Ideally, we would always make a correct decision. In practice, the decision will be based on sample information, some of which may be subject to error. Thus, we can make one of two types of decision errors, as shown in the table below.

		<b><i>True State of Nature</i></b>	
		$H_0$ true ( $H_a$ false)	$H_0$ false ( $H_a$ true)
<b><i>Decision</i></b>	Accept $H_0$	Correct Decision	Type II Error
	Reject $H_0$	Type I Error	Correct Decision

If we reject  $H_0$  when  $H_0$  is true, it is called *Type I error*. The probability of a Type I error is represented by  $\alpha$ . If we accept  $H_0$  when  $H_0$  is false, it is called *Type II error*. The probability of a Type II error is represented  $\beta$ . There is an apparent relationship between  $\alpha$  and  $\beta$ . As  $\alpha$  increases,  $\beta$  decreases, and as  $\alpha$  decreases,  $\beta$  increases. The only way to reduce both  $\alpha$  and  $\beta$  simultaneously is to increase the amount of information available in the sample (i.e. increase the sample size).

The probability of making a Type I error usually is controlled by the researcher, thus it is often used as a measure of the reliability of the conclusion and is called the *significance level* of the test.

We consider our situation of testing for a disease cluster. Given a set of data, we first identify the null and alternative hypothesis with a fixed significance level. Here, the null hypothesis would be “no evidence of disease cluster”, and the alternative would be “evidence of disease cluster.” If we set the significance level very small, then the probability of making a Type I error is very small; i.e., claiming evidence of a cluster when, in fact, there was no cluster (cases randomly distributed). However, we have to pay a higher price of making Type II error, or making a decision that there is no cluster when in fact, there is a cluster. Thus, with this significance level, we can reduce the risk of sending health department people to the region in doubt of clustering, but it also increases the risk of failing to detect the real cluster. Since our goal in this project is to identify the cluster effectively, we must not miss any of the possible clusters in our region. Thus, it would be optimal to set the significance level at the level where the department can afford to send people for further examination as many times as possible, so that, though they may have to submit to a potentially larger number

of “false alarms”, they detect nearly every cluster in the area.

## 3.2 Test Statistics

The null and alternative hypotheses are usually characterized in terms of some parameters in a model that describes the data. In this way,  $H_0$  can be characterized in terms of the model parameters (e.g., “average distance between pairs of cases is constant  $D$ ”), and  $H_a$  can be characterized under different values of the parameters (e.g., “average distance between pairs of cases is not constant”). The “test statistic” allows us to transform the description of  $H_0$  and  $H_a$  from words into numerical values. The test statistic measures how much the data deviate from the null hypothesized value of the parameter. The test statistic is usually standardized so that its sampling distribution, under the null hypothesized value of the parameter, has a well-known form.

For example, if  $H_0$  is “population mean is  $D$ ” and  $H_a$  is “mean is not  $D$ ,”  $\bar{x}$  and  $S_x$  are the sample mean and sample standard distribution of the data, then

$$\text{Standardized test statistic} = (\bar{x} - D)/(S_x/\sqrt{n})$$

which conveniently has (approximately, by the Central Limit Theorem) a standard normal distribution (mean 0, standard distribution 1) if  $H_0$  is true. (If  $H_0$  is not true, then the mean of the data is not  $D$ , and the mean of the standardized test statistic is not zero.) Most of the test statistics in this report will have the form *test statistic* = (*data average - contemplated mean under  $H_0$* )/(*standard deviation under  $H_0$* ), whose distribution under  $H_0$  will be either exactly or approximately normal. When counts are small, the normal approximation may not hold, and other distributions (e.g., Poisson, binomial) are more appropriate. We will see both types of situations in this report.

## 3.3 Rejection Region

After deriving the standardized test statistic, we need to specify the range of possible values, which will be deemed “plausible” or “implausible” under the null hypothesis. In other words, we need to find specific values of the test statistic that will lead us to reject the null hypothesis.

These specific values are known as the *rejection region* for a significance test of level  $\alpha$ . Note that once we decide on  $\alpha$ , the rejection region is defined.

### 3.4 Making a Decision

To complete the test, we make our decision by observing whether the computed value of the test statistic lies within the rejection region. If the computed value falls within the rejection region, we will reject the null hypothesis; otherwise, we do not reject the null hypothesis.

A frequently computed value called the *p-value* assesses the strength of the decision on the hypothesis test. The p-value is the probability under the null hypothesis of observing as extreme a value as the observed test statistic. Computationally, the p-value is the proportion of the distribution of the test statistic under the null hypothesis that lies beyond the observed value of the test statistic. (See Figure ...) If the test is conducted with the significance level  $\alpha$  then a p-value greater than  $\alpha$  is equivalent to failing to reject the null hypothesis using significance level  $\alpha$ . Conversely, a p-value less than  $\alpha$  is equivalent to rejection of the null and evidence in favor of the alternative hypothesis.

### 3.5 Summary of the hypothesis testing procedures

*Step 1:* Specify the null and alternate hypotheses,  $H_0$  and  $H_a$ , and the significance level,  $\alpha$ .

*Step 2:* Determine the test statistics and compute the standardized test statistic using the sample data.

*Step 3:* Determine the rejection region based on the significance level.

*Step 4:* Make the appropriate conclusion by observing whether the computed value of the test statistic lies within the rejection region. If so, reject the null hypothesis; otherwise, do not reject the null hypothesis.

## 3.6 Multiple significance tests and the Bonferroni Correction

When we test a null hypothesis, which is true in fact, with the significance level of  $\alpha$ , the probability that we make the right decision (i.e., correctly fail to reject the null hypothesis) is  $1 - \alpha$ . However, when we conduct the test on  $n$  independent null hypotheses, which are true in fact, with the significance level of  $\alpha$ , the probability that we make the right decision (i.e., fail to reject all the null hypotheses) is  $(1 - \alpha)^n$ . For example, if we conduct 20 tests with null hypotheses all true with  $\alpha = 0.05$ , then the probability we land on the right decision all 20 times is  $(1 - 0.05)^{20} = 0.36$ . Thus the probability of making a wrong decision on at least one of the 20 null hypothesis tests is  $1 - 0.36 = .64$ . That is, the probability of making at least one incorrect decision is greater than the probability of getting all decisions right. In fact, the expected number of wrong decisions on the null hypotheses is  $20 \times 0.05 = 1$ ; i.e., we can expect, with high probability, that we will make at least one error testing. In order to solve this problematic phenomenon, we apply “Bonferroni correction” to the significance level of each test, as follows. Consider the situation where all the null hypotheses are true, and each hypothesis test is conducted at significance level  $\alpha/n$ , instead of  $\alpha$ , where  $n$  is the number of tests. Then, repeating the derivation above, the probability of making an incorrect decision in any one of the null hypotheses is no more than  $n(\alpha/n) = \alpha$  (and possibly less, if the hypotheses are correlated, as they would be if the same data are used in different ways), and the probability of making the right decision on all  $n$  (true) hypotheses is at least  $1 - \alpha$ .

# Chapter 4

## Data and Software

### 4.1 Warning

Notice: The dataset received from the Colorado Health Department was scrambled prior to being distributed to the students involved in this project. This was done to protect the privacy of all individuals whose disease cases are contained in the dataset. The disease codes and location of the cases were both altered, therefore statistical results presented in this paper will differ slightly from results obtained from an unscrambled dataset.

### 4.2 Dataset used in this study

The data from the Colorado Health Department contained 40,819 records, each containing a reported case of a given disease. The records were mostly in the period of 1996 to 2004 (93.7% of them), but the earliest record is from 12/31/1990. Each record has 13 fields, detailed in the table below:

Field Name	A numeric disease code. <b>Altered</b> for privacy reasons. Students were not told what disease each code corresponds to.
Lat	Latitude of the location where the affected patient lives. <b>Altered</b> for privacy reasons.
Long	Longitude of the location where the affected patient lives. <b>Altered</b> for privacy reasons. There were 23,946 unique pairs of latitude and longitude in the dataset.
Tract	The US Census tract where the patient lives.
Pop100	Population for the corresponding census tract.
Cotract	County FIPS code followed by tract code. The tract code is left padded with zeros to ensure a minimum of 6 digits.
Cofips	FIPS code for the county where the patient lives.
cntytrB	Concatenation of fields Cofips, Tract, and Block Group.
Tract1	Same as tract, except for 1,078 of the cases. The meaning of the difference is not known.
Blkgrp	US Census block group where the patient lives. Block groups are defined within the context of a US census tract.
Stdif	Concatenation of the digit 8 (which is the US Census state code for Colorado), plus the field CntyTrB.
Pop2000	Population for the corresponding block group.
Date	Date when the case was reported. Prior to 1998, these dates are not reliable for privacy reasons.

In addition to the full dataset containing the 40,819 cases, a subset was used containing only disease codes 15, 17, 27 and 28. This reduced dataset contained 10,453 records. The fields in each record were exactly the same as the ones in the full dataset.

### 4.3 Software Used

A custom software application was developed to allow students to work with the dataset. The original version of the dataset was split into 3 files. The first file contained columns 1 through 9 for each disease case, the second file contained columns 10 through 12, and the third file contained the dates for each disease case, with five dates per line. A program was written to convert this original dataset into a single comma-separated-value (CSV) file containing all of the data. This CSV file was then used to import the dataset into several applications for analysis and calculation of statistics.

Another program, DataExporter, was developed to export data from the CSV file into a format that can be read by the third-party application ClusterSeer. This custom data export tool allows researchers to pick the disease codes to be exported, as well as which columns should be included in the destination file. It can also output a header that can be read by ClusterSeer.

ClusterSeer is a commercial software package that can be used to calculate cluster-related statistics for certain data sets. We used ClusterSeer version 2.2.0 in this project and found the tool to have extremely limited functionality. One of the serious gaps is the lack of any method to automate the calculation of statistics. This means that calculating cluster-related statistics on an ongoing basis would require significant amount of time from a worker, possibly requiring a full-time or part-time employee. The visualization of test results is very primitive as well, being almost meaningless. Finally, the application is geared towards the calculation of cluster statistics for historical sets of data, rather than towards detection of a possibly ongoing cluster. In general, the tool was found inappropriate for the needs of the Colorado Health Department.

The other major software package used in the project was R, which is a scripting language and environment geared towards statistical applications. R is not a commercial product and is freely distributed from its website. The R language was used by some groups in this project to write implementations of relevant statistical tests. R is a very rich environment and can be used to implement any of the statistical tests described in this paper. The drawbacks of using R is the need to write and maintain custom code to calculate the statistics. This may have significant cost implications for any entity writing its own implementations of tests. On the other hand, R offers enormous flexibility and allows the tests to have exactly the functionality that is needed, including for example the need for automation.

A “package” of R functions for statistical tests for disease clusters can be downloaded from <http://www.r-project.org>.

A special-purpose program called SatScan (<http://www.statscan.org>) was used to calculate Kulldorf’s Spatial Scan Statistic. This program was found to be very useful, but calculates only one statistical test method (see Chapters 6 and 7).

All of the relevant files and tools (except for ClusterSeer, which is commercial and cannot be distributed) were kept in a private web page

that could be accessed by only the students in this course.

# Chapter 5

## Temporal Tests

### 5.1 Introduction

In this chapter, we describe methods of testing for the existence of disease clusters in a relatively short period of time. The data for such tests are one or more time series consisting of counts of disease in time intervals (e.g., number of cases per week or per month). Test statistics rely on one or more series of cases over time. After deciding on whether the data consist of a single or multiple time series, one can choose among several methods. Below we discuss all the available tests for temporal clustering in Cluster Seer: Dat's, Ederer-Myers-Mantel, Empty Cells, Grimson's, Larsen's and the Levin and Kline's modified Cusum tests. All of these methods can be tested for both single and multiple series except for the Ederer-Myers-Mantel, which can be used to detect clusters only within multiple time series. When using count data, all methods assume that the population does not change much over time.

### 5.2 Dat's Method

This method tests for temporal clustering within a single time series or in multiple time series simultaneously, by using counts of cases in consecutive time intervals (a time series) for several areas. An example is the number of cases of measles per week in 10 counties over 6 weeks. This test is more sensitive than the Ederer-Myers-Mantel test

in detecting multiple clusters within a sub-unit and assumes that the population size does not change over time. When the test statistic  $A$  is small, it indicates clustering of cases in one of a few time intervals.[?]

*Data Requirements:*

- Data are counts of cases in each time interval
- Apply the method on only 5-10 time intervals (to avoid seasonal effects)
- Do not use the method when the expected number of cases in each interval is smaller than 2. Therefore, the total number of cases in the time series must be greater than twice the number of time intervals.

If the total number of time intervals exceeds 10, it may be useful to aggregate the data by combining the time intervals. For example, if there are 14 time intervals in the data set, the new data can be obtained by aggregating every 2 intervals. Thus, the new data set has 7 time intervals instead. However, Dat's method applied to a disease such as influenza is likely to detect the winter months as a cluster, so it is important to apply the method to either only short time spans or to diseases that do not exhibit seasonal patterns.

*Analysis:*

$H_0$ : Cases occur at random over the  $t$  time intervals

$H_a$ : Cases do not occur randomly through time

*Notation:*

- $t$  : Number of time intervals
- $n$  : Total number of cases observed over  $t$
- $\frac{n}{t}$ : Number of cases expected in an interval (average number of cases per interval)
- $A$  : The test statistic, equal to the number of time intervals with at least  $\lceil \frac{n}{t} - 0.5 \rceil$  cases.
- $\lceil x \rceil$ : The least integer greater than  $x$ . For example,  $\lceil 1.3 \rceil = 2$ .

Since the cases are distributed randomly across the  $t$  time intervals, the expectation and variance under the null hypothesis are:

$$E(A) = t(0.6 + 0.3d) \text{ and } Var(A) = 0.155E(A)$$

where

$$d = \frac{n}{t} - \left\lceil \frac{n}{t} - 0.5 \right\rceil$$

The formula for  $\text{Var}(A)$  is based on an approximation.

*Significance:*

$A$  is expected to be normally distributed with a mean of  $E(A)$  and variance  $\text{Var}(A)$ . For *single time series*, the  $z$ -score can be used to find the p-value:

$$z = \frac{A - E(A)}{\sqrt{\text{Var}(A)}}$$

For *simultaneous time series* or multiple time series, a chi-square statistic with one degree of freedom is used to find the p-value:

$$x^2 = \frac{\left( \left| \sum_{i=1}^s A_i - \sum_{i=1}^s E(A_i) \right| - 0.5 \right)^2}{\sum_{i=1}^s \text{Var}(A_i)} \quad (5.1)$$

*Note:* Make sure that all of the data requirements are fulfilled before testing. This test should not be used if the data assumptions do not hold. Test statistic  $A$  will be small when cases cluster in one or a few time intervals. On the other hand, a large  $A$  means cluster avoidance (i.e., cases are spread out across the  $t$  time intervals). Usually, it means some of the intervals have slightly more than the expected number of cases. Significant time clustering is indicated by small p-value, e.g., less than 0.05 for a 5% test.

*Recommendations:* Several features about Dat's method caution against its use for clustering analysis. First, the method is biased by changes in population in time. This could be a problem for Colorado, where population in counties such Douglas have increased considerably in the past few years. Second, aggregated data for 11 or more time intervals will not work for seasonal diseases such as influenza. For example, if  $n = 12$  months, and the data are combined in two-month blocks (Jan-Feb, March-April...), Dat's method will detect "clustering" in Jan-Feb versus July-August. Dat's method works only for short data series, 5 to 10 time intervals, to avoid problems with seasonality (e.g., 5 or 10 weeks).

### 5.3 Ederer-Myers-Mantel Method

This method tests for time clustering in multiple time series simultaneously. An example is the number of cases of leukemia in 10 counties by month over 1 year. Unlike Dat's method, it is insensitive to differences in population size over the area. However, it is biased by changes in population size through time. The test statistic is  $m_1$ , the maximum number of cases in a time series. Indication of clustering occurs when  $m_1$  is large. On the contrary,  $m_1$  will be small when cases occur uniformly through time.

*Data Requirements:*

- Data must be counts only (not rates)
- The number of time intervals must be between 2 to 5

As with Dat's method, when there are more than 5 time intervals, we can analyze the subsets or reorganize the data into fewer time intervals.

*Analysis:*

$H_0$ : Cases occur at random in each time series

$H_a$ : Cases do not occur randomly through time, they either cluster or occur uniformly

*Notation:*

- $t$ : Number of time intervals
- $T$ : Number of time intervals in the time series
- $r_i$ : Number of cases in time series  $i$
- $f(r)$ : Frequency, over all time series, of a given number of total cases
- $m_{1i}$ : Largest number of cases in any time interval of time series  $i$

*Significance:*

The data from several time series are used to construct a Chi-squared statistic to test for time clustering in several areas simultaneously. The p-value then can be obtained using:

$$x^2 = \frac{\left( \left| \sum_{i=1}^T m_{1i} - E\left( \sum_{i=1}^T m_{1i} \right) \right| - 0.5 \right)^2}{\sum_{i=1}^T Var(m_{1i})} \quad (5.2)$$

The summations are over the number of time intervals in the time series, where  $\sum_{i=1}^T m_{1i}$  is the sum of the maximum number of cases over all time series.  $E\left(\sum_{i=1}^T m_{1i}\right)$  and  $\sum_{i=1}^T Var(m_{1i})$  are sums of the expectation and variance of  $m_{1i}$  under the null hypothesis.

The calculation for this method can be very time consuming. It involves the calculations for all possible combinations of cases in a time series, and a multinomial calculation for each of the combination. Moreover, this method can be applied only for multiple time series. This can only further complicate the calculations. As  $r$  gets large, computing the distributions for multiple time series will become very time-consuming, even with the computer.[?]

One advantage of the Ederer-Myers-Mantel test is that it is insensitive to differences in population sizes across the areas. However, it is biased by changes in population in time. So it faces the same problem as the Dat's method. This method can take a long time to compute; if the total number of cases in a time series is large, it will require more time.

*Notes:* With Cluster Seer, this test uses only simulated values for data with more than five time intervals. For large values of  $r$ , the functions for  $E(m_{1i})$  and  $Var(M_{1i})$  can be generated by Monte Carlo methods [?]. Time series with 1 or 0 cases are automatically dropped from the analysis because they are assumed to be non-clustered.

*Recommendations:* In table 5.2, the overall chi-square value indicates significance. If we look further into the outputs as well as from the data set, the following counties indicate significant clustering: 13, 31, 35, 41, and 59. In particular, county 41 and 59 have small p-values. The original article by Ederer, Myers, and Mantel states that this test is quite powerful for detecting clusters in samples of sizes around a 200-300, when clustering is extremely intense (e.g., in the 1950's a sudden outbreak of hepatitis).[?]

1	0	0	0	1	0	1	0	0	4	65	0	0	0	0	0	0	0	0	1
5	0	3	0	0	0	1	0	0	4	69	1	2	1	1	0	0	1	0	1
13	0	2	0	0	2	0	0	0	5	75	0	0	0	0	0	0	0	0	1
21	0	0	1	0	0	0	0	0	0	77	0	1	0	0	0	1	0	0	1
29	0	0	0	1	0	0	0	0	0	81	0	0	0	0	1	0	0	0	1
31	0	0	2	1	1	1	0	0	5	83	0	0	0	0	0	0	0	0	1
35	0	0	1	0	0	2	0	0	5	93	0	0	0	0	1	0	0	0	0
41	1	1	1	0	0	0	0	0	10	101	0	1	0	1	0	0	0	1	3
43	0	0	0	0	0	0	0	0	1	103	0	0	0	0	0	0	0	0	1
45	0	0	1	0	0	0	0	0	2	123	1	0	1	1	0	0	0	0	0
59	0	1	0	0	1	2	0	0	11	125	0	0	0	0	0	0	0	0	1

Table 5.1: Excerpt from the data set for disease code 15 from Dec. 31, 1990 to Dec. 31, 1996.

Column 1 denotes the county identification. Column 2 to 10 denotes the date as follow: 19901231 19911231 19921231 19931231 19941231 19951231 19960427 19961210 19961231

Ederer-Myers-Mantel Test \*\*\*\*\*

Number of series = 22

Number of cells per series = 9

Series	Cases	M1	SimuE(M1)	SimuVar(M1)	P-Values
1	6	4	2.09009	0.358609	0.0185525
5	8	4	2.44845	0.393883	0.0938342
13	9	5	2.63063	0.443590	0.00500435
31	10	5	2.87588	0.603816	0.0366089
35	8	5	2.45145	0.400196	0.00120267
41	13	10	3.44545	0.622021	1.73828e-011
45	3	2	1.32132	0.236329	0.713210
59	15	11	3.81181	0.752125	2.56923e-011
69	7	2	2.28428	0.321907	0.616331
77	3	1	1.32633	0.252122	0.515757
81	2	1	1.11612	0.102736	0.717151
101	6	3	2.09209	0.352232	0.491892
123	3	1	1.33333	0.254509	0.508782

Monte Carlo runs used in simulation: 999

Chi-square (1 df) :115.6574275 (using the simulated values)  
Significance : 0.000000 (using the simulated values)

Table 5.2: Above is the exact output from Cluster Seer for the disease code 15 from Dec. 31, 1990 to Dec. 31, 1996 (Ederer-Myers-Mantel Test).

## 5.4 Empty Cells Method

This method tests for time clustering in a single time series or in multiple time series simultaneously. Use this test to detect clusters of rare events and some of the time intervals have 0 cases [?]. When the cases cluster, the test statistic  $E$ , the number of time intervals with 0 cases, will be large.

*Data Requirements:*

- Data must be counts only, no rates
- Should be used to detect clusters for rare events, as some of the time intervals are expected to have 0 cases.

*Analysis:*

$H_0$ : Cases occur randomly through time

$H_a$ : Cases cluster in one or more time periods

*Notation:*

- $A$ : the test statistic, the count of the number of cells with zero cases (empty cells)
- $N$ : Number of cases in a time series
- $t$ : Number of time cells

The following are the expectation and variance under the null hypothesis:

$$E(A) = t \left( \frac{t-1}{t} \right)^N$$

$$E((A)_2) = (t)_2 t^{-N} (t-2)^N$$

$$Var(A) = E(A)(1 - E(A)) + E((A)_2)$$

The notation  $(a)_k$  denotes a falling factorial so that  $(a)_k = a(a-1)\dots(a-k+1)$ . For example,  $(4)_3 = 4 * 3 * 2 = 24$ .

*Significance:*

Important note: Another important criterion is that the number of total cases in a time series should not be too large. When the number of cases,  $N$  is too large,  $E(A) < 1$ . Almost surely:

$$E(A) = t \left( \frac{t-1}{t} \right)^N < 1$$

$$\ln(t) + N[\ln(t-1) - \ln(t)] < 0$$

$$N < \frac{\ln(t)}{\ln(t) - \ln(t-1)}$$

It is assumed that  $N$ , the number of cases must satisfy the above condition to apply the method.

For *single time series*: Under the null hypothesis, we want to determine the probability,  $P$ , of obtaining a number of empty cells greater than or equal to  $E^*$ . The significance of  $A$  is evaluated using the exact p-value:

$$P(E \geq E^*) = (-1)^{E^*} \sum_{k \geq E^*}^{t-1} (-1)^k \binom{k-1}{E^*-1} \binom{t}{k} \left( \frac{t-k}{t} \right)^N$$

Note that  $P(E \geq E^*)$  is evaluated as a one-tailed test.

For *simultaneous time series*: p-values are combined using the Bonferroni approach. A Bonferroni approach is used for multiple tests of significance in order to protect against the Type I error. In research, the more tests we apply in a single study, the higher the probability that we are going to make the Type I error. The true  $\alpha$ -level for the entire study will be inflated. An approximation of how much  $\alpha$  increases as the function of the number of hypotheses one test:

$$\text{overall } \alpha\text{-level} = 1 - (1 - \alpha_0)^k$$

where  $k$  is the number of significance tests being done or contemplated and  $\alpha_0$  is the significant level for each individual test. For example, if we think we are setting  $\alpha$  at 0.05 in our study and we are thinking of testing 20 statistical hypotheses, our actual chance of claiming a significant result when there should not be one is approximately  $1 - (1 - 0.05)^{20} = 0.64$ . Thus, we have about a 64% chance of making a Type I error. The Bonferroni approach is used to control the  $\alpha$  inflation problem, by conducting each test using:

$$\alpha_0 = (\text{desired overall } \alpha) / k$$

Therefore, if we are contemplating testing 20 hypotheses, we would test each one using the new  $\alpha_0$  of  $0.05/20=0.0025$  for each separate significant test.

When at least 20% of the areas have an expected number of empty cells of 5 or more, instead of using the Bonferroni approach, the results can be combined as a continuity-corrected chi-square with one degree of freedom.

$$\chi^2 = \frac{\left( \left| \sum_{i=1}^t E_i - \sum_{i=1}^t E(E_i) \right| - 0.5 \right)^2}{\sum_{i=1}^t Var(E_i)} \quad (5.3)$$

Here the  $i$  subscript indicates a statistic for the  $i^{th}$  time series.  $E$  is the sum of the number of empty cells, and  $E(E)$  and  $Var(E)$  are the mean and variance of  $E$ .

*Note:* The Empty Cells method can be used only with count data, not rates. Also, some of the intervals in the time series must have 0 counts (Empty Cells). This method is designed for relatively rare data. The method will not work appropriately if the expectation of the number of empty cells is too small, or if the total cases in a time series is too large [?]. When cases cluster the test statistic will be large. But when equal numbers of cases occur in all the cells,  $E$  will be smaller than its expectation and the test statistic will be small. Clustering is indicated by small p-value or a large Chi-square value. That is, when fewer than 80% of the series used had empty cell counts of 5 or more, a Bonferroni p-value is calculated for significant clustering. When more than 80% of the series used had empty cell counts of 5 or more, we can calculate continuity-corrected chi-square for significant clustering.

Illustration:

Empty Cells Method \*\*\*\*\*

Series	E	E(E)	Var(E)	Upper-tail p-value
1	6	4.43943	0.670047	0.089570
5	6	3.50770	0.845960	0.011523
13	6	3.11795	0.895812	0.003983
31	4	2.77152	0.923162	0.215342

35	6	3.50770	0.845960	0.011523
41	5	1.94652	0.902013	0.003065
45	7	6.32099	0.242646	0.308642
59	5	1.53799	0.832788	0.000628
69	3	3.94616	0.771103	0.962065
77	6	6.32099	0.242646	1.000000
101	5	4.43943	0.670047	0.459432
123	6	6.32099	0.242646	1.000000

Because more than 80% of the series used had empty cell counts of 5 or more, we can calculate a continuity-corrected chi-square for significant clustering:

Chi-square: 32.95408094  
p-value: 9.449857563e-009

Table 5.3: *An example of a Empty Cells test outputted by Cluster Seer, using the same data from Table 5.1.*

*Recommendations:* This method should not be used on any data that is not considered rare. [?] It will not work well if there are not enough zero cases in a time series. From the table outputs, chi-square value indicates significant clustering. In particular, if we look at the p-values for counties 13, 41, and 59, they specify cases clustering. It makes sense if we compare the results with the Ederer-Myers-Mantel test. It seems to work well. However, when the test was run for a few different times with different data sets, it turns out that even with the most obvious clustering period, this test would fail to detect clustering because that particular period has a large number of cases occurred. In particular, consider the following example:

1	0	18	11	0	0	2	0	0	0	0	2	3	0	0	0	0
4	0	0	0	2	0	5	0	0	0	0	0	6	0	0	0	0
7	5	1	0	0	0	8	0	0	0	0	0	9	0	0	0	0
10	0	6	0	1	0	11	0	2	0	1	0	12	0	35	1	0
13	1	28	38	0	2	14	0	0	0	0	0	15	0	0	0	0
16	0	0	0	0	0	17	0	1	0	0	0	18	1	0	0	0

Table 5.4: *An example of how the Empty Cells method can not be applied. The first column denotes the county id and the other columns*

denote time intervals.

The first column denotes the 18 counties, and the subsequent columns denote the cases that occur in that county at a particular time interval. That is similar to the first table. Below are the outputs from Cluster Seer:

Empty Cells Method \*\*\*\*\*

Series	E	E(E)	Var(E)	Upper-tail p-value
7	3	1.31072	0.525853	0.040000
10	3	1.04858	0.508936	0.016192
11	3	2.56000	0.326400	0.520000

Because fewer than 80% of the series used had empty cell counts// of 5 or more, we calculate a Bonferroni p-value for significant// clustering:

Bonferroni p-value: 0.048576

Table 5.5: *The outputs for table 5.4 using Cluster Seer.*

Notice the data in counties 1, 12, and 13 indicate clusters, but the method does not. The reason is that they have large numbers of cases. From the Cluster Seer outputs below, county 7 and 10, which have small but moderate number of cases are identified as clustered. Thus, this method can not detect clustering for diseases that are not considered rare.

## 5.5 Grimson's Method

This method detects clustering of cases in space, time, and space-time. We need to have a data subset of labeled high risk cases to use this

method. When data are rates we have several ways to decide if a time interval is high-risk. The first one is when the rates are significantly large relative to a reference population. Another one is when the rates are in the upper 5% of the rates among the time intervals. We also can use external events as the risk criteria. For example, if an exposure occurred during specific time intervals, we declare these intervals to be the high risk time intervals.

*Analysis:*

$H_0$ : “High risk” items arise at random (no clustering).

$H_a$ : High risk items tend to be adjacent (clustering).

*Notation:*

- $N_t$  : the number of time intervals.
- $n$  : the number of high risk time intervals.
- $y$  : the average number of adjacencies per time interval
- $Var(y)$ : the variance in the number of adjacencies.
- $A$  : the number of pairs of “high risk” objects that are adjacent.

$H_0$ : The objects have been labeled at random. Under this hypothesis the number of adjacencies among the labeled cells is expected to be:

$$E(A) = \frac{yn(n-1)}{2(N_t-1)}$$

The variance of  $A$  has two components, the regularity component ( $RC$ ) and the variability component ( $VC$ ). The regularity component is:

$$RC = E(A) \left( 1 + \frac{2(y-1)(n-2)}{N_t-2} + \frac{(N_t y - 4y + 2)(n-2)(n-3)}{2(N_t-2)(N_t-3)} - E(A) \right)$$

The variability component is:

$$VC = Var(y) \left[ \frac{n(n-1)(n-2)}{N_t(N_t-1)} \left( 1 - \frac{(n-3)}{N_t-2} \right) \right]$$

The variance of  $A$  is

$$Var(A) = RC + VC$$

*Significance:* We evaluate the significance of  $A$  using the Poisson or Normal distribution. The first assumes  $A$  is sampled from a Poisson distribution with a mean given by  $E(A)$ . The second assumes

that  $\frac{[A-E(A)]}{\sqrt{Var(A)}}$  is approximately Normal distribution with mean of 0 and variance 1.0. Both approaches yield a one-tailed test describing the probability, under the null hypothesis.

Whether to use the Poisson distribution or the normal approximation depends on the proportion of the variance,  $Var(A)$ , contributed by the variability component,  $VC$ . This is  $VC/Var(A)$ . Grimson (1991) offers the following guidance: Use the Poisson distribution when  $VC/Var(A)$  is small,  $< 0.20$  Use the normal approximation when  $VC/Var(A)$  is large,  $\geq 0.20$ . Below is an example of the method calculated by Cluster Ser:

Date	Number of cases	Adjacencies	Risk:1 case/day
8/1/1999	0	1	0
8/2/1999	0	2	0
8/3/1999	3	2	1
8/4/1999	1	2	1
8/5/1999	0	2	0
8/6/1999	2	2	1
8/7/1999	0	2	0
8/8/1999	0	2	0
8/9/1999	0	2	0
8/10/1999	0	2	0
8/11/1999	2	2	1
8/12/1999	0	2	0
8/13/1999	8	2	1
8/14/1999	0	2	0
8/15/1999	0	2	0
8/16/1999	1	2	1
8/17/1999	2	2	1
8/18/1999	4	2	1
8/19/1999	1	2	1
8/20/1999	0	2	0
8/21/1999	0	2	0
8/22/1999	0	2	0
8/23/1999	1	2	1
8/24/1999	0	2	0
8/25/1999	0	2	0
8/26/1999	3	2	1
8/27/1999	0	2	0
8/28/1999	0	2	0
8/29/1999	0	2	0
8/30/1999	2	2	1
8/31/1999	1	1	1

Table 5.6: *Table of disease cases for the month of August, 1999. The first two columns are data inputs for Cluster Seer.*

Grimson's Method \*\*\*\*\*

Inputs: Number of Objects = 31

Number of labelled cells = 13

Average number of borders per cell = 1.935483871

Sample variance of the number of borders = 0.062366

Number of pairs of adjacent labelled cells = 5

Outputs:

RC	=	1.774044894
VC	=	0.07907885911
EA	=	5.032258065
Var(A)	=	1.853123753
z-score	=	-0.0236966032

Significance:

normal	=	0.5094526922
poisson	=	0.5651485772
VC/Var(A)	=	0.04267327478

Grimson (1991) suggests that one use the Poisson approach when  $VC/Var(A)$  is small, and the normal approach when  $VC/Var(A)$  is large. One rule of thumb is to use the Poisson approach when  $VC/Var(A) < 0.20$ ; otherwise, use the normal approach.

Table 5.7: *Output from Cluster Seer for disease cases for the data above.*

## 5.6 Larsen's Method

This method tests for dispersion of cases about a central time period, for a disease that is assumed to be rare. It may be used for time clustering in a single time series or in multiple time series. We do not recommend this method because it can't distinguish multiple clusters from the uniform distribution (no clustering).

*Analysis:*

$H_0$ : Cases occur randomly throughout the  $t$  time periods.

$H_a$ : Cases cluster about a single point in time.

*Notation:*

- $t$  : the total number of time intervals

- $m$  : the number of time periods with at least one case
- $y_i$  : the time assigned to the  $i^{th}$  cell in which a case occurred
- $r + 1$  : the index of the ‘central most’ time cell that contained a case,  $r = \lceil m/2 \rceil$ . The operation  $\lceil \alpha \rceil$  is the least integer greater than  $x$ .
- $K$  : the test statistic, measures the tendency of time periods with at least one case to form a single cluster in time. It is

$$K = \sum_{i=1}^m |y_i - y_{r+1}|$$

The expectation and variance of  $K$  under the null hypothesis are

$$E(K) = \frac{(t+1) \lceil \frac{m}{2} \rceil \lceil \frac{m+1}{2} \rceil}{m+1}$$

$$Var(K) = \frac{r(t+1)(t-m)((m+1)^2 - 2r^2 - \delta(m))}{12(2 \lceil \frac{m+1}{2} \rceil + 1)^2}$$

Here  $\delta(m)$  is  $r - 2$  when  $m$  is odd, and  $\delta(m) = 2r - 1$  when  $m$  is even. The test statistic  $K$  can be expressed as a  $z$ -score which is expected, under the null hypothesis of a random allocation of occupied cells across the time series, to be normally distributed with a mean of 0 and unit variance:

$$z = \frac{K - E(K)}{\sqrt{Var(K)}}$$

The distribution of  $z$  is approximately  $N(0, 1)$ .

When occupied time intervals form a unimodal cluster,  $K$  will be smaller than its expectation and the  $z$  score will be less than 0. A uniform distribution of occupied time intervals through time, such as ‘01010101’ will cause  $K$  to be larger than its expectation, and the  $z$  score will be greater than 0.

*Simultaneous time series:* When the data consist of several time series the  $K$  statistics from each time series can be combined into an overall  $z$ -score as

$$z_G = \frac{\sum_{i=1}^S K_i - \sum_{i=1}^S E(K_i)}{\sqrt{\sum_{i=1}^S Var(K_i)}}$$

For an overall departure from the expected values across all time series simultaneously, we use the grand  $z$  score tests. The individual  $z$  scores

test for unimodal clustering within each time series. We must examine the individual  $z$  scores before concluding whether a significant grand  $z_G$  score is due to unimodal clustering in all of the time series, or to some other combination of temporal pattern across time series.

*Notes:* When the time series is shorter than 10 intervals, the normality assumption doesn't hold. For short time series, consider using smaller time intervals or perhaps collecting data from additional time periods. Larsen's method requires two or more of the time intervals to have cases. Time series with fewer than 2 occupied intervals are excluded from an analysis. Also, each time series must have at least 1 unoccupied cell. Counts are required and the method is biased by changes in population size through time. We can not use Larsen's method with rates. The grand  $z$ -score is not biased by differences in population size across time series.

## 5.7 Levin-Kline CuSum Method

This method is based on the CuSum statistic for detecting trends or changes in level of sequences. It was proposed initially for purposes of monitoring production processes. Measurements are made on parts, which, ideally, all take the same value. If the measurements start to drift, or a change in the production process causes a shift in the mean of the measurements, it is important to detect that shift as quickly as possible, before too many defective parts are produced. In the context of disease clusters, the "measurements" are "number of cases in a given time interval" (e.g., one day, or one week, or one month); the average number of cases typically seen in a given time interval is denoted by  $\lambda_0$ . For example, CDPHE might expect to see no more than about 30 cases of pertussis per month for the entire state of Colorado; here,  $\lambda_0 = 30$ . The observed number of cases per month may fluctuate about 30, say, between 20 and 40; if a subtle change in the environment were to cause a gradual shift or increase in this number over time, then the CuSum statistic is helpful for detecting slight changes over time.

We illustrate the computation of this statistic using the data between January 1999 and April 2004 for disease 28. (For details of the computation and the theory associated with this statistic, see B. Levin and J. Kline, "The CuSum Test of Homogeneity with an application

in spontaneous abortion epidemiology,” *Statistics in Medicine* 4: 469–488, 1985.) First, the number of cases for each interval are recorded. For disease 28, the numbers of cases by month are given in the following table:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1999	21	15	26	17	17	18	18	28	29	32	23	38
2000	40	31	37	33	32	18	28	27	70	39	59	39
2001	53	38	21	8	11	9	21	41	24	18	48	65
2002	26	28	24	34	26	19	26	49	28	59	57	47
2003	28	35	19	37	35	16	25	23	24	28	43	26
2004	26	42	57	55								

Table 5.8: *Disease cases from January 1999 to April 2004.*

Let  $Y_t$  denote the number of cases in month  $t$ , where here  $t = 1, \dots, n = 64$ . Denote the typical number of cases, or baseline risk, as  $\lambda_0$ ; e.g., for the data above,  $\lambda_0$  might be 30 cases per month. We first need to determine a limit beyond which an increase in this typical number of cases might be deemed large enough to consider intervention measures. Typical increases might be 20%, 33%, 50%. Here we illustrate with “effect size”  $\omega = 1.33$  (33%), or 40 cases per month. From  $\lambda_0$  and  $\omega$ , we calculate a reference value  $r$  determined as follows:

$$r = \lambda_0(\omega - 1) / \log_e(\omega) \quad (5.4)$$

where  $\log_e(x)$  denotes the natural logarithm of  $x$ . For  $\lambda_0 = 30$  and  $\omega = 1.33$ ,  $r = 34.72$ .

The CuSum statistic is computed iteratively as follows.

- (a) Let  $W_0 = 0$ .
- (b) Each subsequent value of  $W$ ,  $W_t$ ,  $t = 1, 2, \dots, n$ , is computed as the maximum of 0 and  $(W_{t-1} + Y_t - r)$ ; that is, if  $(W_{t-1} + Y_t - r)$  ever goes negative, then set  $W_t$  to zero; otherwise, set  $W_t$  equal to this value  $(W_{t-1} + Y_t - r)$ .
- (c) Find the maximum value of  $W_t$  for all  $t = 1, \dots, n$ . This maximum is  $W_{max}$ , the value of the CuSum statistic.

Note that  $W_{max}$  depends upon the reference value  $r$ , which depends upon the chosen baseline risk  $\lambda_0$  and the effect size  $\omega$ . Calculations for the 64 months of disease 28 counts are shown in Table 5.9.

Month	Count	Count-34.72	$W_t$	Month	Count	Count-34.72	$W_t$
1	21	-13.72	0.00	33	24	-10.72	0.00
2	15	-19.72	0.00	34	18	-16.72	0.00
3	26	-8.72	0.00	35	48	13.28	13.28
4	17	-17.72	0.00	36	65	30.28	43.56
5	17	-17.72	0.00	37	26	-8.72	34.84
6	18	-16.72	0.00	38	28	-6.72	28.12
7	18	-16.72	0.00	39	24	-10.72	17.40
8	28	-6.72	0.00	40	34	-0.72	16.68
9	29	-5.72	0.00	41	26	-8.72	7.96
10	32	-2.72	0.00	42	19	-15.72	0.00
11	23	-11.72	0.00	43	26	-8.72	0.00
12	38	3.28	3.28	44	49	14.28	14.28
13	40	5.28	8.56	45	28	-6.72	7.56
14	31	-3.72	4.84	46	59	24.28	31.84
15	37	2.28	7.12	47	57	22.28	54.12
16	33	-1.72	5.40	48	47	12.28	66.40
17	32	-2.72	2.68	49	28	-6.72	59.68
18	18	-16.72	0.00	50	35	0.28	59.96
19	28	-6.72	0.00	51	19	-15.72	44.24
20	27	-7.72	0.00	52	37	2.28	46.52
21	70	35.28	35.28	53	35	0.28	46.80
22	39	4.28	39.56	54	16	-18.72	28.08
23	59	24.28	63.84	55	25	-9.72	18.36
24	39	4.28	68.12	56	23	-11.72	6.64
25	53	18.28	86.40	57	24	-10.72	0.00
26	38	3.28	89.68	58	28	-6.72	0.00
27	21	-13.72	75.96	59	43	8.28	8.28
28	8	-26.72	49.24	60	26	-8.72	0.00
29	11	-23.72	25.52	61	26	-8.72	0.00
30	9	-25.72	0.00	62	42	7.28	7.28
31	21	-13.72	0.00	63	57	22.28	29.56
32	41	6.28	6.28	64	55	20.28	49.84

Table 5.9: *Cases of Disease 28 between January 1999 and April 2004*

The maximum value of  $W_t$  occurs at month 26 (February 2001), 89.68, surrounded by high values at months 23, 24, 25, and 27 (November 2000 to March 2001). The next highest set of values occurs around month 48 (December 2002), 66.40, surrounded by high values at months 47, 49, 50 (November 2002 to February 2003). The third highest set of

values occurs at month 64 (April 2004), 49.84.

To determine whether these values (89.68, 66.40, 49.84) are large enough to be considered “statistically significant,” we generate 10,000 series of 64 random Poisson counts having mean 30. For each random series, we calculate  $W_{max}$  exactly as above. Among these 10,000 trials, the average  $W_{max}$  was 16.36; only 7 trials returned values of  $W_{max}$  greater than 49.84 (the largest of the 10,000  $W_{max}$  values was 60.64). Because the observed  $W_{max}$  for the disease 28 series was 89.68, larger than any of the values from these 10,000 trial, the significance of that time cluster around January 2001 is significant at 0.0001. The second highest value, 66.40, is also larger than any of the maxima from the 10,000 simulations, so it also has a p-value of 0.0001. The third highest value, 49.84 around November-December 2002, has significance  $7/10,000 = 0.0007$ . A plot of the  $W_t$  series and the three identified maxima of the series is shown in figure 5.1.

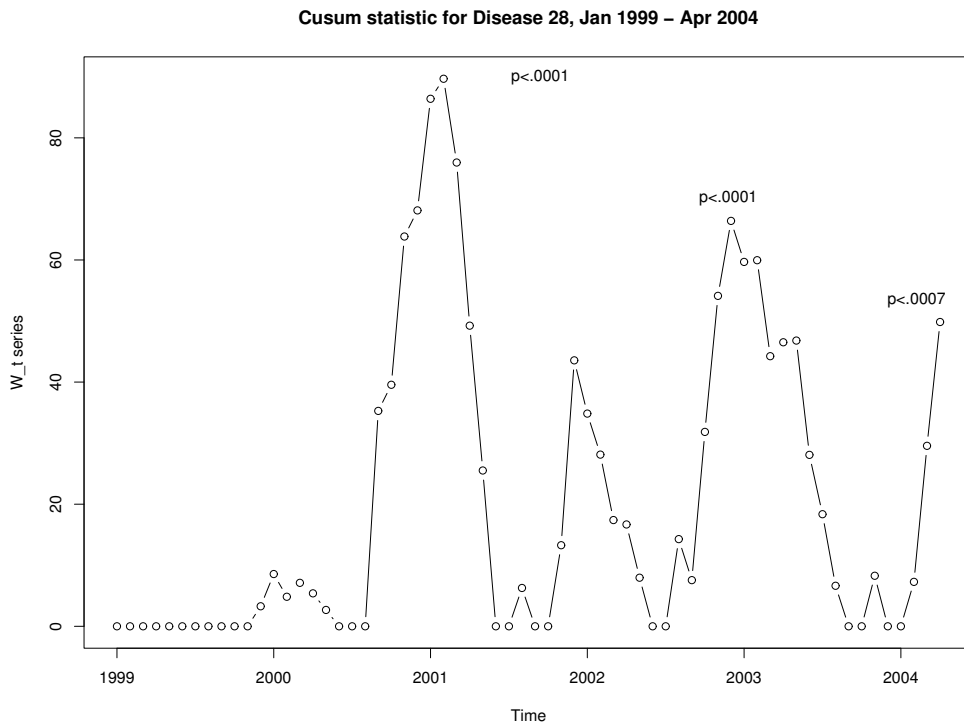


Figure 5.1: *CuSum* statistics for Disease 28, Jan 1999-Apr 2004.

### 5.7.1 Shewhart control charts

For detecting huge, sudden changes in the number of cases of a disease that is not expected to show seasonal variation, a simple “Shewhart control chart” is more effective. In this type of chart, one plots the observed number of cases per unit interval (e.g., per month) as a function of time, along with a dotted horizontal line corresponding to the expected number of cases (e.g., for pertussis, 30 cases), and a solid line for an “upper limit” of  $\lambda_0 + 3\sqrt{\lambda_0}$  (e.g, for pertussis,  $30 + 3(5.5) = 46.5$ ). Months for which this limit is exceeded are flagged. An example is shown in Figure 2, using the data from January 1999 to April 2004 for disease 28 shown in Table 10. The plot shows several months when the upper limit of 46.5 cases was exceeded: September 2000, November 2000, January 2001, December 2001, August to December 2002 (except for September), and March and April 2004. Again it is important to emphasize that this chart is appropriate only for diseases for which no seasonal component is expected.

	1999	2000	2001	2002	2003	2004	Median
Jan	21	40	53	26	28	26	27
Feb	15	31	38	28	35	42	33
Mar	26	37	21	24	19	57	25
Apr	17	33	8	34	37	55	33
May	17	32	11	26	35		26
Jun	18	18	9	19	16		18
Jul	18	28	21	26	25		25
Aug	28	27	41	49	23		28
Sep	29	70	24	28	24		28
Oct	32	39	18	59	28		32
Nov	23	59	48	57	43		48
Dec	38	39	65	47	26		39

Table 5.10: *Cases of Disease 28 between January 1999 and April 2004*

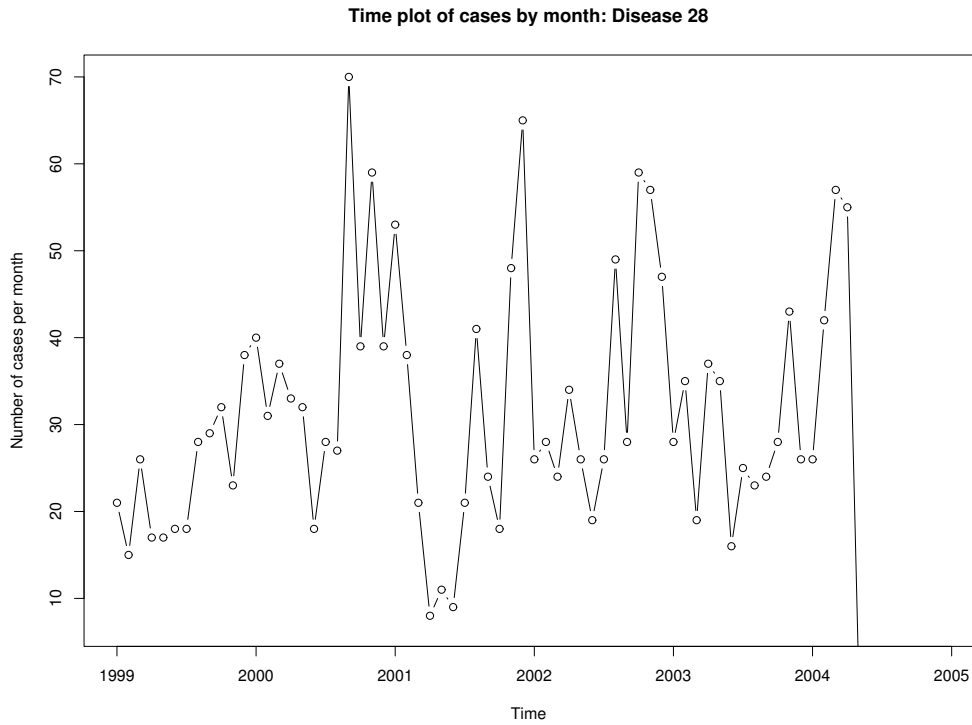


Figure 5.2: *Time plot of cases by month: Disease 28.*

If a seasonal component is expected, then the data should be adjusted for the seasonal effect. One simple way of accomplishing this adjustment for monthly data is to subtract a typical January effect from all January months, and likewise for the remaining 11 months. An example is shown for disease 28. In each row, the median is subtracted, yielding the following “seasonally-adjusted” data:

	1999	2000	2001	2002	2003	2004
Jan	-6	13	26	-1	1	-1
Feb	-18	-2	5	-5	2	9
Mar	1	12	-4	-1	-6	32
Apr	-16	0	-25	1	4	22
May	-9	6	-15	0	9	NA
Jun	0	0	-9	1	-2	NA
Jul	-7	3	-4	1	0	NA
Aug	0	-1	13	21	-5	NA
Sep	1	42	-4	0	-4	NA
Oct	0	7	-14	27	-4	NA
Nov	-25	11	0	9	-5	NA
Dec	-1	0	26	8	-13	NA

In this series, one expects to see zero cases after having accounted for the seasonal effect (the median of these 64 numbers is zero), and most observations are within three standard deviations (or about twice the interquartile range; here, twice 10.5, or 21). A plot of these seasonally-adjusted counts of cases is shown in Figure 5.3.

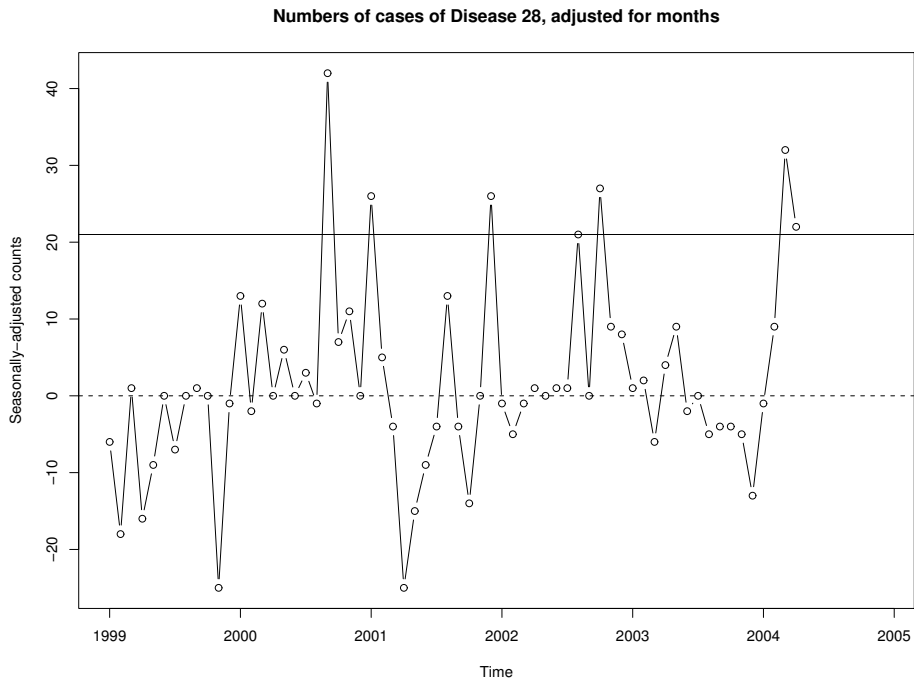


Figure 5.3: *Number of cases of Disease 28, adjusted for months.*

This figure shows that counts for September 2000, January 2001, December 2001, August and October 2002, and March and April 2004, are unusually high, even after having adjusted for a possible monthly trend. (Karen Kafadar)

## 5.8 Sudden increase in number of cases: MMWR's "Figure 1"

Stroup et al. (1993) recommended a method for detecting aberrations in public health surveillance data that has come to be known as "Figure 1" in the *Morbidity and Mortality Weekly Report*, made available to epidemiologists, clinicians, and other public health professionals in a timely manner. Although the tables of the *Morbidity and Mortality Weekly Report* provide important information, the volume of data and the need for ease of interpretation encourage the development of a graphic display to highlight unusually high or low numbers of reported cases.

An analytical and graphical method was developed to achieve the following objectives: (1) to depict in a single comprehensible graph weekly reports of approximately 20 disease totals that can be compared with past results; and (2) to highlight for further analysis the results most likely to indicate changes in long-term trends or epidemics. These objectives were formulated to reflect most recent behavior in as short a time period as possible for weekly publication, but long enough to provide stable results. To facilitate comprehension, the same is used method for all diseases.

The analytic method currently used as figure 1 in the *Morbidity and Mortality Weekly Report* (figure 1), called the Morbidity and Mortality Weekly Report Current/Past Experience Graph of the Centers for Disease Control and Prevention, compares the number of reported cases in the current 4-week period for a given health event with historical data from the preceding 5 years (6, 7). Numbers of cases in the current 4-week period are listed to facilitate interpretation of instability caused by small numbers.

The choice of 4 weeks as the "current period" was based on evidence of weekly fluctuation in disease reporting that is usually due to irregular reporting than to disease incidence. The use of a 5-year history

achieves the objective of applying the same model for all conditions depicted. This is particularly helpful since some health events were made notifiable only recently (e.g., acquired immunodeficiency syndrome and legionellosis). In addition, modeling of reported influenza incidence has shown that more accurate forecasts are based on more recent data (8). To increase the historical sample size and to account for any seasonal effect, the baseline is taken to be the average of the reported number of cases for the preceding 4-week period, the corresponding 4-week period, and the following 4-week period, for the previous 5 years. This yields 15 correlated observations, referred to as the historical observations for baseline.

The deviation from unity of the ratio of the current total to the historical average is indicative of a departure from past patterns. We plot this ratio on a logarithmic scale so that an  $n$ -fold increase projects to the right the same distance as an  $n$ -fold decrease projects to the left, and no change from past patterns (1:1) produces a bar of zero length (9). To distinguish the conditions that may require further investigation, the hatching on the bars begins at a point based on the mean and standard deviation of the historical observations (6). (Historical limits of the ratio of current reports to the historical mean are calculated as  $1 \pm 2$  times the standard deviation divided by the mean, where the mean and the standard deviation are calculated from the 15 historical 4-week periods.)

Because surveillance data are reported sequentially in time, they may not satisfy the assumptions necessary for usual time series analyses. For example, the number of measles cases reported to the Centers for Disease Control and Prevention in a given 4-week period over 5 years, 1985-1989, is highly correlated from period to period (Pearson product-moment correlation = 0.86). The problem is particularly apparent with incidence data for which the numbers of reported cases are subject to seasonal effects and reporting delays. The method used to set the hatching point for the Current/Past Experience Graph may be affected by the correlation of reported health events over time, and little is known about the empirical performance of the method in the presence of such correlation.

Since this study was conducted, the method has remained useful in national surveillance. Recent increases beyond historical limits in reporting of aseptic meningitis reflected increased disease activity primarily in the northeastern United States. Increases in animal rabies reports were due to increased reporting of raccoon rabies in mid-Atlantic and

northeastern States (unpublished data). Although not applicable to all types of surveillance questions, the method is useful for identifying conditions that require further investigation and for providing sensible solutions from imperfect data to facilitate public health action. For more information on this method, see Stroup et al. (1993).

Reference Stroup, D.F., Wharton, M., Kafadar, K., Dean, A.G.: An evaluation of a method for detecting aberrations in public health surveillance data, *American Journal of Epidemiology* 137: 373-380, 1993.

## 5.9 Conclusion

The Dat's, Ederer-Myers-Mantel, Empty Cells, Grimson's, Larsen's and Cusum are six common temporal cluster detection tests out there that are implemented by Cluster Seer.

- Most of these methods require the data to be counts, not rates.
- We do not recommend Dat's method because it is biased by changes in population in time. The aggregated data proposal will not work because some newly aggregated will be more clustered than others.
- Dat's method works only for short data series, 5 to 10 time intervals. So if we have a large set of data, we must adjust it for seasonality.
- The Ederer-Myers-Mantel method is a little less strict compared to Dat's. The method is insensitive to differences in population size across the areas. However, it is still sensitive to changes in population over time. Cluster Seer recommends its use when there are two to five time intervals in the series. In such situations, where the data has more than five time intervals, the values will be simulated by the software using the Monte Carlo method. In addition, the Ederer-Myers-Mantel method can be used only for multiple time series.
- The Empty Cells test can be applied for rare events only, and some of the time intervals must have 0 counts.
- Grimson's method can be used with counts or rates.
- Larsen's method only can be used with counts (not rates), is biased by changes in population size through time and can not distinguish multiple clusters from a uniform distribution (no clustering).

- Cusum method is based on the Cusum statistic for detecting trends or changes in level of sequences. It was proposed initially for purposes of monitoring production processes. We recommend this method because if a subtle change in the environment were to cause a gradual shift, then the Cusum statistic is helpful for detecting slight changes over time, and appears to be the most sensitive and responsive to clustering in time.

# Chapter 6

## Spatial Test

### 6.1 Introduction

The spatial methods are statistics based cluster detection methods designed to investigate clusters over a specified region. Spatial cluster detection methods are categorized into several different classifications. Global (also referred to as General) methods scan the assigned area for clustering (e.g. entire state of Colorado), and have no preconception of the location of possible clustering. Global methods scan the entire region to identify statistically significant clusters, which may not have occurred by chance. Local methods are classified as either General Local or Focused Local; both are designed to test for clustering at a specified location. The General Local methods test observations (e.g. occurrence of a disease) at geographical locations (which may have been identified prior by a Global test), by examining the neighbors of an observation (within a specified distance) for statistical significance. For example, if a case of West Nile Virus may occur two miles west of Niwot, other nearby occurrences of the disease are considered when testing if clustering exists in the area or if the occurrence is a random event. Note that the clustering of the disease may be deemed significant by the General Local test but the specific location of the disease around Niwot may not be known. Unlike the General Local tests, the Local Focused tests are designed to detect clustering around specific source which is proposed to increase the risk of a disease. For example, a Focused test might be performed around Rocky Flats to test for an increase in certain types of cancer. The (alternative) hypothesis in this case would be “clustering exists around Rocky Flats.”

The spatial methods examined in this analysis are the Global and General Local, assuming no preconception of the location of the clustering. If West Nile Virus were hypothesized to occur at a specific location, then a Focused test would be appropriate. However, the intent of this study is to scan for clusters of West Nile Virus or other diseases, with no preconception of where clustering might occur. In this analysis, the data are aggregated at the county and the census tract levels, using a centroid of the county or census tract as the location of all disease cases in that region.

When testing using either the Global or Local methods, the null hypothesis is the same, that there is "no evidence of clustering". The alternative hypothesis is that the cases of disease that do not occur randomly in space. Statistical significance is measured by a p-value which, when small (e.g.  $p < 0.10$  or  $p < 0.05$ ) indicates rejection of the null hypothesis. According to the epidemiology of West Nile Virus, which does not behave like a vector borne disease, each occurrence is assumed to be independent, and the risk of being contagious is negligible. The spatial tests included in this analysis specify that the risk of contracting the disease is due to a population risk, and/or the proximity of the disease. The spatial models are designed to look at past data and then provide inferences on possible clustering. These methods are not designed to predict future clustering. Finally, this chapter discusses only tests for spatial clustering; tests for spatial and temporal clustering are discussed in a later section of this report.

## 6.2 Methods

A review of the following methods will be discussed in detail in the following sections. The models to be covered are:

- (a) Cuzick and Edwards [Global]
- (b) Besag and Newell [Global]
- (c) Moran(I-Statistic) [Global and General Local]
- (d) Geary(I-Statistic) [Global]

The aforementioned spatial methods implement various techniques to scan the area of study. Nearest neighbor techniques (e.g., Cuzick and Edwards) assess significance to the nearest neighbors around a case of the disease. Other methods are based on drawing a centroid (or another

specified shape, trapezoids, rectangles, etc) to scan for clusters (e.g., the methods: Besag and Newell ,Moran's I) use circles. The effectiveness of each method depends on the assumptions used to specify the test. These assumptions include: relevant population at risk, calculation of the risk of contracting the disease, characteristics of the data , and the relevant time periods of data to include in the analysis. The following sections detail implementation of the individual methods, including model specification and the underlying assumptions.

## 6.3 Besag and Newell

### 6.3.1 Overview

The method proposed by Besag and Newell is a specialized version of the Geographical Analysis Machine (GAM) proposed by Openshaw)[?]. GAM methods scan a spatial area (e.g. Colorado) centered at specified points(county or census tract centroids). At each centroid the GAM then generates a radius from the point and tests the region defined by the radius for significance (e.g., clustering). The GAM then tests all radii at each point for the entire region (e.g., Colorado). This process is computationally intensive. Besag and Newell proposed testing only the regions with radii which contain only specific numbers of occurrences of the disease. For example the Besag and Newell statistic could specify scan regions with 10 cases, and base the calculation of statistical significance using the total population in these regions. This method can be applied as either a Global and Focused spatial test. This report implemented the Global test to examine occurrences of West Nile Virus in Colorado. The Focused test, although not used, is also discussed. The Global method is designed as a screening device to detect possible clustering in areas which may merit further investigation. Additional analysis may require the use of statistical methods or alternate scientific methods. This test is designed to work best when testing of a small number of cases relative to the population at risk. The cases are assumed to be independent. Since West Nile Virus is not considered an infectious disease, this assumption may be appropriate for this disease (where the cases are roughly independent of the density of the human population).

### 6.3.2 Applications

The GAM tests were applied originally to examine childhood leukemia, which was the first major attempt to discuss clusters of a rare disease. The original research by Besag and Newell also applied their test to study the occurrences of childhood leukemia. Other applications include testing for leukemia in children over a region, and testing other for clustering of other rare diseases, and clustering of rare non-epidemiological events. Like West Nile Virus, the leukemia discussed in the studies is not a contagious disease, which makes the application of Besag and Newell this test to West Nile Virus reasonable.

### 6.3.3 Method Specification

The effectiveness of the Besag and Newell method is dependent on numerous assumptions. The assumptions to be specified are: what time period(s) of data to use, risk of disease, level of significance of the test, and the number of cases that defines a cluster. First, the question of which time periods of data to use is important, since different factors could affect the data from different time periods in the data set. Also, this method is designed to detect clusters in historical data, and makes no inferences or predictions of future clustering.

Calculating the risk of contracting the disease is not a trivial exercise. This method assumes that all people have the same risk of contracting the disease. So a probability could be calculated by taking the total cases with the total population. The situation could arise where the whole area has a significantly different number of cases in different time periods. Analysis must be done to determine the probability of contracting the disease. Besag and Newell discuss alternate techniques for calculating this rate, such as Cliff and Ord (1981).

Next, the user must choose a number of cases (e.g., West Nile Virus). The method identifies appropriate regions for significance. Different numbers of cases will affect which clusters appear statistically significant. Also, the number of cases can be changed when going from the Global method to a General Local method, to further assess whether a cluster is valid. Finally a  $p$ -value must be chosen as a cutoff for the test; if it is too strict then relevant clusters may not be detected, and if the  $p$ -value is too large, insignificant clustering may be reported.

### 6.3.4 Methodology

#### Regions

The regions are built around each centroid (county or census tract), including those tracts with no reported cases. At each centroid the closest centroid is added to the region until the number of cases specified is reached. Only those regions with the specified number of cases or greater will be considered by the method.

#### Global Spatial Testing

Tests of Global spatial testing are designed as a screening device to detect possible clusters. The null hypothesis states that the observed total number of cases  $y$  is distributed entirely at random among the population at risk. The General Focused method tests the null hypothesis states that each person contracts the disease with probability  $p$  which is a specification defined by the user calculated from the historical data. Each test examines only the regions with the specified number of cases or greater. No assumption is made as to the exact location of the disease clustering.

- $k$  = minimum number of cases required for the test statistic for each region
- $y_j$  = number of cases in a region  $j$ .
- $t_j$  = population of a region.
- $A_o$  label assigned to the region in which the cluster occurs
- $A_i = \{1,2,3,\dots\}$ : regions determined by the increasing distances of the centroids from  $A_o$ .

Calculate  $D_i = \left(\sum_{j=0}^i y_j\right) - 1 =$  total number of cases in  $A_i$  (accumulated).

Calculate the accumulated numbers of population at risk  $u_0 \leq u_1 \leq \dots$  such that

$$u_i = \left(\sum_{j=0}^i t_j\right) - 1$$

The number of cases in a region follows a Poisson distribution with a common rate.

Let  $M = \min\{i : D_i \geq k\}$  i.e. zone  $A_M$  but not  $A_0, \dots, A_{M-1}$ , contains at least  $k$  other cases.

The null hypothesis is the hyper-geometric probability that  $s$  individuals among  $u_m$ , is approximated by the Poisson term ( note  $u_m * p = \lambda$ , and  $p = y/t$ )

$$\frac{e^{-\lambda} \lambda^s}{s!}$$

It follows that

$$Pr(M \leq m) = 1 - Pr(M > m)$$

$$Pr(M \leq m) = 1 - \sum_{s=0}^m \frac{e^{-\lambda} \lambda^s}{s!}$$

### 6.3.5 Discussion

Overall the Besag and Newell method is intended as a screening test to detect clusters. This test may detect numerous regions where further analysis may be required to determine if a cluster occurred. Specification of the assumptions is key for this test. The data could be grouped into smaller and larger historical groups, to study where historically the clustering has occurred in the past; the rate of the disease could be changed; or the number of cases used to specify a cluster could be changed. It is recommended that this test be run multiple times, to better detect the presence of clustering. (One needs to be mindful of the problems associated with multiple testing.) This method is more effective in detection of clustering for a low number of occurrences of the disease; see the ‘Recommendations’ section of this report for the results the Besag and Newell method applied to the Colorado data.

## 6.4 Cuzick-Edwards Method

### 6.4.1 Overview

One major factor to consider is that the Cuzick-Edwards method is more effective with diseases with longer incubation periods. This method

has been used with leukemia cases in New Zealand [?]. The definition of incubation period is subjective. The literature includes cases for Leukemia and other rather long term diseases. This brings to question whether this method would work well with West Nile Virus. Cuzick-Edwards statistic would certainly not be appropriate for diseases such as measles or influenza, as those infections are far too fast moving. This study found a high rate of clustering in Leukemia and was a very appropriate use employment of the Cuzick-Edwards method. One advantage of the Cuzick-Edwards method is that it does not need specific information on populations such as gender and age [?], because it considers nearest neighbors and does not rely simply on distance data. However, the disadvantage is that the method requires knowledge of all negative cases as well as all the positives, which requires knowing how many people live near a positive case and, of those, how many are infected. This information may not be readily available.

### 6.4.2 Example

The major example is from the Dockerty paper [?] and focuses on leukemia and lymphoma. They found 748 cases in the years studied. Only six hundred of these cases were used because birth records were not available for all the cases. In this example the Cuzick-Edwards test is modified with upper and lower statistical bounds using the method of Jacquez [?]:

“The method of Cuzick and Edwards is particularly suitable for use in populations that have an uneven geographical distribution, as in New Zealand.”

The major advantage of this method is that it is robust in unevenly populated areas. In Dockerty’s study, controls were matched for age, gender, etc also, this method is not recommended for finding small clusters by Dockerty [?]. This method is a good global method for long term studies because of its lack of sensitivities to differences in population distributions and to its nearest neighbor approach, which does not require intense calculations of exact distance.

### 6.4.3 Assumptions and inputs

The assumptions of this method are:

- (a) the total population is known
- (b) the number and locations of positive cases of the disease are known
- (c) the location of the cases and reasonably accurate of the distances between cases and populations.

One possible modification to the application of this method is that we might use census tract centroids calculated upon the weighted values of populations, because census tracts are small and we can better estimate where people actually live, thereby eliminating the need to calculate distances between all persons. For rare diseases this will work well and has great possibilities; for more common diseases, a more complicated modeling approach must be followed as a rate would be necessary for each centroid and for the Cuzick-Edwards method. We would have to empirically find a rate, based on epidemiology. The epidemiology would indicate a minimum threshold rate above which the count would be increased, rather than a simple one or zero criteria with which to work. The parsimony and simplicity of the inputs of this method and lack of need for exact distances are the arguments in favor of using Cuzick-Edwards. The problem generally speaking, with simple models is that they tend not to provide as much information. This is a simple model, so falls victim to providing less information, but is not as sensitive to information that might mislead more complicated models. This test can use data on the gender and age of population, but is not necessary[?].

#### 6.4.4 Methodology

The method involves a test statistic, an expected value for the null hypothesis, and a  $z$  score which is calculated for the preceding values to see if clusters are statistically significant [?].

The Cuzick-Edwards clustering method uses a nearest neighbor approach to look for clusters. If an occurrence of a case within a given distance from another case is found, it is registered in the summation that is part of the machinery of the method [?]. It is important to note that distance is not in terms of physical distance with the nearest neighbor method, but rather it is based on a count of the number of cases among the  $K$  nearest neighbors as a measure of distance.

The test statistic is composed as follows:

First this method uses a binary delta function which can be either 1 or zero to identify cases. The function is one if there is a case and zero if

not. This is then multiplied by a similar  $d$  function which is one if there is another case within the specified distance and zero otherwise. The nearest neighbor distance could be changed depending upon population densities as a weight of more less populous areas [?].

- (a)  $\delta_i = 1$  if the  $i^{th}$  case is positive otherwise  $\delta = 0$   
 $d_i^k = 1$  if the  $k^{th}$  nearest neighbor is also a positive, otherwise  
 $d = 0$  .

The test statistic is calculated as:

$$T_k = \sum_{i=1}^N \delta_i d_i^k$$

The expectation value is calculated using the sample size and the population size. This is a straight forward calculation from the number of cases and the total population:

$$E(T_k) = pkN$$

where  $p$  is defined as follows:

$$p = \frac{N_{positive}}{N} \left( \frac{N_{positive} - 1}{N - 1} \right)$$

and

$N_{positive}$  refers to cases that are positive and  $N$  is the total population.

Finally the  $z$  score is calculated from the test statistic, its expected value under the null hypothesis, and the square root of the variances.

$$z = \frac{T_k - E(T_k)}{\sqrt{Var(T_k)}}$$

The rather complicated function for variance in this case may be found in The value of  $Z$  refers to a cumulative normal distribution function.

### 6.4.5 Discussion

This method would hardly be usable in a short term incubation disease, but might find use in longer duration of incubation diseases. For CD-PHE the Cuzick-Edwards method would be indicated for identifying clusters in long-term diseases. It might be inappropriate for seasonal diseases such as West Nile Virus, and be especially ineffective for even shorter term diseases such as influenza, measles, and whooping cough. This model is also very useful for looking at the state wide cases for a historical outlook, but not as useful on local levels and certainly not as useful for projections into the future.

## 6.5 Spatial Autocorrelation

Tobler's first law stated that:

"everything is related to everything else, but near things are more related than far things"

[?]. In spatial autocorrelation methods, nearby data are more important than far away data. Tobler's first law is an important concept to measure spatial autocorrelation. According to Upton and Fingleton [?], the director of studies in Land Economy in St John's College states that, "...spatial autocorrelation as a property that mapped data possess whenever it exhibits an organized pattern." In other words, if there is a pattern or the same type of measurement occurs in different locations, then spatial autocorrelation will find that pattern. Moran I, Geary C, and local Moran I are well-known as spatial autocorrelation methods. These methods basically share a common structure; all measure the significance of a case based on distances to nearby cases.

The hypotheses for spatial autocorrelation tests are:

$H_0$  : Presence of a case is independent of the locations of other cases,

$H_1$  : Presence of a case is related to locations of other case.

## 6.6 Global Moran I

### 6.6.1 Application

In 1998, the department of Ecology, Evolution, and Neurology of State University of New York at Stony Brook and the EMMES Corporation in Potomac, Maryland applied the spatial autocorrelation methods, Moran I and Geary C to forty cancer mortality distributions in Western Europe. The test analyzed three hundred and fifty five different areas by using spatial autocorrelation methods. The authors also applied the test to different subsets of the population such as males only, females only, and the total population. It is common for statisticians to use a combination of spatial autocorrelation tests such as Moran I or/and Geary C, local Moran I and/or local Geary C. An example is the test on the geographical distribution of male and female lung cancer cases in Nassau, Queens, and Suffolk counties in New York [?].

### 6.6.2 Method Specification

According to Sawada, a member of department of Geography in University of Ottawa, Moran I is a method to measure the spatial autocorrelation of ordinal, interval or ratio data [?]. In other words, Moran I is a method to measure the randomness of data (Sawada 1).

The most important component in spatial autocorrelation methods is the matrix of spatial weights. There are numerous techniques to calculate the spatial weight, but two well-known techniques for specifying the matrices for spatial autocorrelation are the binary matrix, and the distance matrix. Binary matrix assigns value of one for regions which sharing a border with the assigned location, and a value of zero for regions not sharing a border. On the other hand, the distance matrix is calculated by a function of distance between the two locations  $i$  and  $j$ . Different methods use different distance functions such as  $\frac{1}{w_{ij}}$ , or  $\frac{1}{w_{ij}^2}$ , or  $\frac{1}{(1-w_{ij})^2}$ . Different kinds of distance matrices can be used to test different diseases such as the binary matrix for a disease that does not spread out, and a distance matrix for a disease contagious.

The goal of Moran I test is to examine the Global clustering. The Moran I value ranges from negative one (-1) spatial autocorrelation to positive one (+1) spatial autocorrelation. If the value is positive, the observed values are similar or the cluster exists. However, if the value

of Moran is approach to 0 or negative, this means that there is no clustering in the study region.

### 6.6.3 Methodology

Moran I statistic:

$$I(d) = \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_{i=1}^n \sum_{j=1}^n W_{ij}}$$

where,

$n$  is number of observations

$x_i$  is the value at location  $i$

$x_j$  is the value at location  $j$

$\bar{x}$  is the mean given by

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$W_{ij}$  is the spatial weight as given in model specification.

$$S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

is the observed variance of population.

## 6.7 Global Geary C

### 6.7.1 Method Specification

The other popular method in spatial autocorrelation is Geary C. Geary C and Moran I are similar spatial autocorrelation method and hence many authors use either one of them. The aforementioned example from State University of New York is a good example of where both

methods are used. The authors used Moran I to analyze the incidences cancer in Western Europe. They also used Geary C to compare the results to Moran I, because these methods contain basically similar structure. The difference in Moran I and Geary C is that value of Moran I ranges from negative one (-1) to positive one (1). On the other hand, value of the Geary C statistic ranges from zero (0) to two (2), where a value less than one (1) indicates that the data are similar and more clustered than those greater than zero (0), that are further apart. The other differences between Moran I and Geary C is that Geary C describes the clustering at a more local level than Moran I.

### 6.7.2 Methodology

The formula for Geary C statistic is,

$$I(d) = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n W_{ij} (x_i - \bar{x})(x_j - \bar{x})}{2 \sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n \sum_{j=1}^n W_{ij}}$$

where the notation is the same as that of the Moran I. Global spatial methods test for the presence of clustering, but they do not show the location of the clusters. To detect specific clusters, we must use local spatial methods.

## 6.8 Local Moran I

### 6.8.1 Application

In 1998, Local Moran I was used to examine the hepatitis cluster in all the counties in California [?]. The test is calculated based on the average population from 1995 to 1997 for all counties. The results detected clustering in the northwest corner of California. The test was also run in Global Moran I and Geary C before. However, the results for Global Moran I and Geary C only show the existence of clustering but not location. While local Moran I use the same data to detect five counties where are clustering exists.

## 6.8.2 Method Specification

In 1995, Anselin develop a local indicator of spatial association (LISA statistic) Local Moran I based on Moran I. From geographical analysis the properties of LISA are

1. The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;
2. The sum of LISAs for all observations is proportional to a global indicator of spatial association.[?]

The second property is very important to distinguish between LISA methods and other local spatial methods. Since local Moran I is derived from global Moran I, they both have the same spatial weight  $W_{ij}$ . Also, the results of Local Moran I are very similar to Global Moran I. High positive spatial autocorrelation value refers to similar data and clustering. If the value of Local Moran I is less than or approach to zero (0) from both positive and negative, this indicates no clustering exists and data are dissimilar.

## 6.8.3 Methodology

The Local Moran I statistic is:

$$I_i(d) = (x_i - \bar{x}) \sum_j^n W_{ij}(x_j - \bar{x})$$

where the notation is the same as that used for global Moran I (section 6.3).

## 6.8.4 Discussion

Global Moran I provides non-specific evidence of clustering, while local Moran I indicates the location of cluster. Local Moran I is recommended before sending resources to area of positive clustering. Population in a distance study can be a problem for this test, so some authors recommend using zip code for a distance. The other problem is that the test require a best understanding of the distances in studying. Choosing a good distance to make the test will receive the

best result. Moreover, local Moran I was used to detect the West Nile Virus before in Chicago. Carmen Tedesco (a member of department of Geography in University of Illinois), Connie Austin (a member of Illinois Department of Public health) and others applied local Moran I method on 680 cases of West Nile Virus in and around Chicago in 2002 [?]. Since local Moran I have already been useful for West Nile Virus in Chicago, it will also be useful for Colorado. .

## **6.9 Recommendations**

### **6.9.1 Besag and Newell**

#### **Inputs**

Centroid locations of county and census tract, risk of contracting disease, population, number of disease that defines a region,  $p$  value, data grouped by year (2000-2004)

#### **Implementation**

R used with DCluster (package) which was quite easy, testing was done using county and census tract data

#### **Results**

Results are sensitvie to inputs, however, the model did do a good job of finding clusters quickly on a global level. No results for 2004 since we only had data through April, 2004 and the rate was based upon the rate for the entire year for other statistical run; no statistically significant clusters were found for 2004. Besag and Newell worked well on diseases w/low counts(e.g. disease number 15). Not as effective for other diseases (e.g. 17,28,87) since those had too many cases.

### **6.9.2 Cuzick-Edwards**

#### **Inputs**

General location of positive cases and centroids of census tracts.

## Implementation

Method used centroids of census tracts as an aggregate of the information took positive cases into account with the test statistic. This was implemented as far as possible in R.

## Results

There is computational difficulty involved in computing this method's variance therefore the method could not be calculated fully for statistical significance.

### 6.9.3 Moran Local Global and Geary C

The popular spatial autocorrelation methods include the Global Moran I, Global Geary C, and Local Moran I. All these methods are dependent upon the spatial weight which requires a good geographical information base, while the test does not require a lot of information on population. Local Moran I is recommended because the test can focus on the location of individual points which a Global statistic may fail to detect. Also, Local Moran I is very popular, it can be calculated by many software packages like ClusterSeer, especially some freeware do a good job with the calculation such as R-project, Rookcase [?]

# Chapter 7

## Spatio-Temporal Clustering

### 7.1 Introduction on Spatio-Temporal Clustering

Spatio-temporal clustering refers to the greater density of occurrences of a phenomenon in certain places at certain times. For example, a disease with localized outbreaks, where cases arise close together, both geographically and in time, will exhibit spatio-temporal clustering. Aggregating the data over time will obscure the pattern if most portions of the region under study eventually experience an outbreak. Likewise, aggregation over space to get a time sequence of incident totals will obscure the pattern if outbreaks of comparable size are fairly evenly spaced in time. Tests for space-time clustering may be used to determine if the distribution of a disease is consistent with a proposed etiology, or to suggest possible etiologies. For example, a contagious disease requiring close contact for infection might show spatio-temporal clustering as the disease spreads from infected individuals. Detection of localized emerging outbreaks is also a problem in analyzing data for spatio-temporal clustering.

These uses lead to several broad categories of statistical tests of spatio-temporal clustering, corresponding to similar categories for spatial clustering. Some tests are designed to determine if the distribution of a phenomenon in the region and time under study shows a general pattern inconsistent with spatio-temporal uniformity. These tests detect space-time interaction. By contrast, another class of tests identifies the

most marked clusters in the data, and assesses them for significance. A third type of test attempts to answer the question of whether a cluster exists around a preselected subregion in time and space. In consultation with individuals active in statistical analysis in public health, in search of past/current journal articles and their bibliographies, and in examination of the ClusterSeer software techniques, the Knox test and the Kulldorff spatio-temporal scan stood out as effective, tested methods for addressing the first two questions. Focused tests for space-time clustering are less established, though existing spatial focused cluster tests could be adapted.

In its original form, the Knox test recommended here falls into the first category. It tests whether the number of points (e.g., occurrences of a disease) that are close in space and also close in time is consistent with the number of points that are close in both space and time if space and time are independent. Evidence of clustering arises if the number of points close in both space and time is significantly higher than would be expected under independence. Other tests for space-time interaction include the Jacquez k-nearest neighbor test [?] and the Mantel test [?]. The Knox test has the advantage that the selection of parameters is simple and intuitive.

The Kulldorf scan statistic is an especially effective method for cluster identification. The method scans the data for geographic subregions that, during some time interval, have unusually large numbers of cases. The method simulates the distribution of the scan statistic under the null hypothesis of no clustering and then compares the observed value of the statistic with that distribution to determine the whether the most marked cluster is consistent with spatio-temporal uniformity. These properties place the Kulldorff method in the second category, that of cluster identification. Raubertas proposed another such test [?]. The Kulldorff test is widely used, in part due to the availability of software to perform the analysis.

Tests of whether certain points represent foci of clusters tend to be designed for clustering in space alone. Waller and Gotway (2004) present the framework of score tests of the hypothesis that clusters exist around pre-selected foci in space.[?] The weight functions used to represent exposure to the foci in such tests could be generalized to be functions of time.

Cluster detection may be done retrospectively, answering the question, “When and where were the clusters, if any, in the fully developed data set?” Prospective cluster detection applies to the circumstance that the data set is generated as an ongoing process, with data for recent cases added as the cases are reported. With each inclusion of new data, prospective analysis addresses the question “Is a new cluster developing?” The detection of emerging outbreaks most naturally uses prospective methods. Prospective variants of both Knox’s method and Kulldorff’s method are described below.

Several caveats apply to the methods detailed here. Both require determination of parameters by the researcher. In order for the significance calculations of the tests to be meaningful, these parameters must be determined from general considerations of the phenomenon being examined, rather than the specific appearance of the data. Multiple tests under a range of values of the parameters also complicate the analysis of the significance.

Infectious diseases and environmental risk factors may not produce detectable clustering or clusters if the location and time of exposure are not strongly related to the location and time of the reported case, due, for example, to a population that is very mobile on the scale of the latency period of the disease.

Both methods require substantial supplemental analysis for use with null hypotheses differing from the hypothesis of spatio-temporal uniformity. For example, each test may be biased by seasonal clustering, or changes in population not reflected in the input to the method.

Finally, the Knox tests and the most common versions of the Kulldorff spatio-temporal scan statistic do not test for purely spatial or purely temporal clustering. These tests should be used in conjunction with the spatial tests and temporal tests described in the previous chapters.

## 7.2 Kulldorff Space-Time Scan Statistic

### 7.2.1 Introduction

This method identifies the most significant cluster of a particular shape in space-time.

In the context of identifying disease clusters, the methods apply to group-level data. For a fixed geographic subregion of the geographic region under study, there are multiple measurements over time, giving the number of new cases of the disease and a value for the population at risk in that subregion during the interval since the last measurement. The location (centroid) of the measurement is given by a pair of spatial coordinates for the center (by some definition) of the subregion and a time coordinate for the interval. Certain types of subsets of the centroids are considered potential clusters. These subsets are called *zones*. Typically a zone consists of all centroids with space coordinates within a particular geographic boundary, and time coordinates within a particular interval. For example, if the geographic bases are disks, the zone boundaries are right circular cylinders.

The method identifies the zone showing the strongest evidence of representing a high density cluster. The scan statistic is based on a maximum likelihood ratio for each potential cluster that expresses how much more likely the observed density is under the hypothesis of clustering than under the hypothesis of uniformity. The value of the retrospective scan statistic for a data set is the maximum value of these maximum likelihood ratios over the collection of zones. The prospective scan statistic takes the maximum with respect to zones with time intervals terminating at the time of the most recent measurement.

The significance of the observed value for the scan statistic is based on a Monte Carlo simulation. A low p-value provides evidence that the cluster is more extreme than can easily be explained by chance.

The methods are not designed to detect overall clustering.

The more detailed discussion below largely follows Kulldorff (1997)[?].

## 7.2.2 Notation

$G$ : The set of centroids of space-time regions in the study.

$Z$ : A variable denoting a zone. The subsets of  $G$  that may be used as zones are described more fully in Section 2.4.

$\mathcal{Z}$ : The set of zones being examined.

$N$ : A spatio-temporal random process. For each subset  $A$  of the centroids,  $N(A)$  is the random count, e.g. of cases of a disease, assigned to  $A$ . For example, if the centroids represent counties, the  $N(A)$  is the total of the numbers of cases in the counties whose centroids lie in  $A$ .

$n_A$ : The observed count associated with  $A$ , i.e. the actual number of cases associated with all centroids in  $A$ .

$\mu(A)$ : A weight indicating the proportion of counts expected in  $A$ . The value of  $\mu(A)$  is often the total population at risk in the regions with centroids in  $A$ .

## 7.2.3 Models

The null and alternative hypotheses for the scan statistics may be based on a Bernoulli model or a Poisson model. In either model, the notation above applies.

For the Bernoulli model, each subset of centroids,  $A$ , is assigned a value  $\mu(A)$  corresponding to the total number of units in the at-risk population in each centroid in  $A$ . Because a centroid represents a geographic region during a time interval, the at-risk population is in person-time units, such as person-years. The null hypothesis states that the probability that any unit represents a case of the disease is a Bernoulli trial with probability  $p$ , where  $p$  is constant for all the regions. The alternative hypothesis says that there is one zone  $Z$  such that the probability that any unit in  $Z$  is a case is a Bernoulli trial with probability  $p$ , while the probability that any unit outside  $Z$  (denoted  $Z^c$ , the complement of  $Z$ ) is a case is a Bernoulli trial with probability  $q$ , with  $q < p$ . To summarize:

$H_0$ :  $N(A) \sim \text{Bin}(\mu(A), p)$  for all sets  $A$

$H_1$ :  $N(A) \sim \text{Bin}(\mu(A), p)$  for all sets  $A \subset Z$ , and  $N(A) \sim \text{Bin}(\mu(A), q)$  for all sets  $A \subset Z^c$ , with  $p > q$ .

Under the Poisson model, the measure  $\mu(A)$  may be the size of the population-at-risk for the region  $A$ , but may be chosen more generally to reflect the expected proportion of cases for that region determined by other criteria. For example,  $\mu(A)$  may be calculated from a regression analysis of covariates for the disease.

$H_0$ :  $N(A) \sim \text{Poisson}(p\mu(A))$  for all sets  $A$

$H_1$ : There is exactly one zone  $Z$  and  $p > q$  for which

$$N(A) \sim \text{Poisson}(p\mu(A \cap Z) + q\mu(A \cap Z^C)) \text{ for all sets } A$$

The Bernoulli model and the Poisson model will yield similar results for a rare disease in a large population. Kulldorff (1997) suggests using the Bernoulli model for binary counts, such as cases and non-cases.[?] This model describes the situation in which each individual has a certain risk of contracting the disease, independent of others. This is not true for contagious diseases, of course. Even when the alternative hypothesis is an inaccurate model for the disease, it may yield a higher maximum likelihood than the null hypothesis, keeping the method effective.

Cluster Seer implements the Poisson model. This is best suited to situations in which one unit can contribute multiple counts, such as number of episodes of a recurrent condition.

## 7.2.4 Zones

A zone is a potential cluster, generally defined as the set of all the centroids falling within a geometric figure, typically a cylinder with space coordinates within a geometric figure in space and the time coordinate within a particular interval. The following are possible sets of geometric figures for the spatial bases of the cylinders:

- All circles centered at each of the subregional centers and containing at most half the population
- All circles centered on foci of a grid, with a possible upper limit on size
- All rectangles of a fixed size and shape

- Ellipses centered at centers or foci

Circles alone will not adequately address clustering along linear features such as highways, waterways, powerlines, or mountain ranges.

More customized zones may be chosen for a particular region, if that choice is based on general features of the region and the health events under study. For example, one may choose to examine school districts as zones, if one suspects the disease is spread among school children. The Monte Carlo significance procedure will uniformly use these fixed zones to produce a valid result. By contrast, noting an apparent cluster in the data, gerrymandering a zone to take in that cluster, and using that zone in the subsequent Monte Carlo testing will not produce a valid result.

If customizing zone shapes to the data, the researcher must have a replicable process for determining the customized shapes, and apply that process to each simulated distribution in the Monte Carlo computation to retain the validity of the significance. Some work has been done using simulated annealing to allow for zones consisting of arbitrary collections of contiguous subregions.[?]

For spatio-temporal analysis, Kulldorff has used cylinders consisting of circular regions in space over varying time intervals. Nothing prevents the use of cylinders with other bases, such as ellipses or rectangles.[?]

## 7.2.5 Computation of the Scan Statistic

Recall  $n_Z$  denotes the number of cases in zone  $Z$ , and  $n_G$  denotes the total number of cases in the study. Let  $n_{Z^c}$  denote the number of cases in the entire study that are not in zone  $Z$ .

For the Bernoulli model, the likelihood of observing  $n_Z$  and  $n_G$  under the alternative hypothesis is

$$L(Z, p, q) = p^{n_Z} (1 - p)^{\mu(Z) - n_Z} q^{n_{Z^c}} (1 - q)^{\mu(Z^c) - n_{Z^c}}$$

This is maximized over  $p$  and  $q$  by  $p = \frac{n_Z}{\mu(Z)}$  and  $q = \frac{n_{Z^c}}{\mu(Z^c)}$  if  $\frac{n_Z}{\mu(Z)} > \frac{n_{Z^c}}{\mu(Z^c)}$ , and by  $p = q = \frac{n_G}{\mu(G)}$  otherwise. (Just use calculus to find the extremum over the set of  $p$ 's and  $q$ 's with  $p > q$ .)

The likelihood under the null hypothesis is

$$L(Z, p) = p^{n_G} (1 - p)^{\mu(G) - n_G}$$

This is maximized by  $p = \frac{n_G}{\mu(G)}$ .

For each zone under consideration, compute the ratio of those likelihoods:

$$\lambda_Z = \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(1 - \frac{n_Z}{\mu(Z)}\right)^{\mu(Z) - n_Z} \left(\frac{n_{Z^c}}{\mu(Z^c)}\right)^{n_{Z^c}} \left(1 - \frac{n_{Z^c}}{\mu(Z^c)}\right)^{\mu(Z^c) - n_{Z^c}}}{\left(\frac{n_G}{\mu(G)}\right)^{n_G} \left(1 - \frac{n_G}{\mu(G)}\right)^{\mu(G) - n_G}}$$

if  $\frac{n_Z}{\mu(Z)} > \frac{n_{Z^c}}{\mu(Z^c)}$ , 1 otherwise.

Then the scan statistic  $\lambda$  is given by the maximum of all the  $\lambda_Z$ 's:

$$\lambda = \max_{Z \in \mathcal{Z}} (\lambda_Z)$$

where  $\mathcal{Z}$  is the set of all zones.

For the Poisson model, the likelihood corresponding to the alternative hypothesis is

$$L(Z, p, q) = \frac{e^{-p\mu(Z)} (p\mu(Z))^{n_Z}}{n_Z!} \frac{e^{-q\mu(Z^c)} q\mu(Z^c)^{n_{Z^c}}}{(n_{Z^c})!}$$

again maximized over  $p$  and  $q$  by  $p = \frac{n_Z}{\mu(Z)}$  and  $q = \frac{n_{Z^c}}{\mu(Z^c)}$  if  $\frac{n_Z}{\mu(Z)} > \frac{n_{Z^c}}{\mu(Z^c)}$ , and by  $p = q = \frac{n_G}{\mu(G)}$  otherwise.

The likelihood under the null hypothesis is

$$L(Z, p) = \frac{e^{-p\mu(G)} (p\mu(G))^{n_G}}{(n_Z)!(n_{Z^c})!}$$

This is maximized by  $p = \frac{n_G}{\mu(G)}$ .

For each zone under consideration, compute ratio of the maximum likelihoods:

$$\lambda_Z = \frac{\left(\frac{n_Z}{\mu(Z)}\right)^{n_Z} \left(\frac{n_{Z^c}}{\mu(Z^c)}\right)^{n_{Z^c}}}{\left(\frac{n_G}{\mu(G)}\right)^{n_G}}$$

if  $\frac{n_Z}{\mu(Z)} > \frac{n_{Z^c}}{\mu(Z^c)}$ , 1 otherwise.

Again the scan statistic  $\lambda$  is given by the maximum of all the  $\lambda_Z$ 's:

$$\lambda = \max_{Z \in \mathcal{Z}} (\lambda_Z)$$

In the case of a prospective analysis, the same computations apply, but  $\mathcal{Z}$  includes only zones with time intervals extending to the time of the most recent observations.

Though zone definitions may involve infinitely many geometric figures, there will be only finitely many distinct zones because there are finitely many centroids. Thus the maximum of the  $\lambda_Z$ 's may be determined exhaustively.

Large values of  $\lambda$  indicate clustering of cases in the zone, while  $\lambda \approx 1$  indicates that the number of cases of disease in the zone is about average.

### 7.2.6 Monte Carlo simulation

To determine the significance of the retrospective  $\lambda$  for the data, compute  $\lambda$  using the same  $\mathcal{Z}$  for 999 simulations of the distribution of cases under the null hypothesis. (The default number of simulations in ClusterSeer is  $N_s = 999$ .) Estimate the upper tail probability by  $1/(N_s + 1)$  (e.g.,  $1/1000$  if  $N_s = 999$ ) times the number of simulated  $\lambda$ -values greater than or equal to the  $\lambda$  for the data. To determine the significance of the prospective  $\lambda$  for the data, compute  $\lambda$  for each simulated data set using the corresponding  $\mathcal{Z}$  but removing the restriction that the time interval extend to the most recent data.

To simulate a sample from the null distribution under the Bernoulli model, given the total number of observed cases  $n$ , associate each centroid  $x$  with an interval of  $\mu(x)$  numbers in  $(0, \mu(G))$ . Draw a subset of size  $n$  from the distribution in which each subset of  $(0, \mu(G))$  is equally likely. Set  $N(x)$  equal to the number of the elements lying in the interval corresponding to  $x$ . Simulation if  $n$  is unspecified is also simple. Sample  $(Bin\mu(x), p)$  for  $p = \frac{n_G}{\mu(G)}$  to get  $N(x)$ .

To simulate a sample from the null distribution under the Poisson model, given the total number of cases  $n$ , use a multinomial distribution with  $m$  categories, one corresponding to each centroid. The probability of the category for centroid  $x$  is  $\frac{\mu(x)}{\mu(G)}$ . If the number of cases is unspecified, sample  $Poisson(p\mu(G))$  to choose a value.

Software called SaTScan, downloadable from <http://www.satscan.org/>, implements various versions of the Kulldorff spatial scan method. The software supports retrospective and prospective tests, and can perform adjustments for multiple categorical covariates.

### 7.2.7 Example

This simple example using an artificial data set illustrates these analyses.

Suppose there are three regions, North, Center, and East, with the geographic centroids, populations, and case counts indicated below. Here, the populations are constant. In general, one provides the best estimate of the at-risk population at the time of the each observation.

<i>Region</i>	<i>Centroid</i>	<i>Population</i>	<i># cases (2003)</i>	<i># cases (2004)</i>
North	(0,1)	10,000	20	5
Center	(0,0)	10,000	25	10
East	(1,0)	20,000	10	5

Consider the zone consisting of Center in 2003. Here,  $n_z = 25$ ,  $\mu(Z) = 10,000py$ ,  $n_{Z^c} = 50$ ,  $\mu(Z^c) = 70,000py$ ,  $n_G = 75$ , and  $\mu(G) = 80,000py$  (40,000 for each of 2003 and 2004), where  $py$  is a person-year. Substituting these values into the Poisson model's formula for  $\lambda_Z$  and taking the natural log gives  $\lambda_Z \approx 11$ . (The natural log helps keep the values in a range practical for computation.)

For comparison, consider the zone consisting of North and Center in 2003. Here,  $n_z = 45$ ,  $\mu(Z) = 20,000py$ ,  $n_{Z^c} = 30$ ,  $\mu(Z^c) = 60,000py$ ,  $n_G = 75$ , and  $\mu(G) = 80,000py$ . Substituting these values into the Poisson model's formula for  $\lambda_Z$  and taking the natural log gives  $\lambda_Z \approx 21$ .

In fact, this last  $\lambda_Z$  is the maximum, if we take  $Z$  to consist of right circular cylinders whose base includes no more than half the population, and whose height includes no more than half the time period of the study. This definition of  $Z$  is the default in ClusterSeer and SaTScan. Note that the subset consisting of Center in 2003 and 2004 is not a zone by this definition.

The Monte Carlo p-value for  $\lambda = 21$  is .001, as estimated with 999 simulated data sets in SaTScan. The SaTScan analysis notes that the maximum  $\lambda$  from the simulated data was about 4. Thus any  $\lambda$  greater than 4 would have .001 as its p-value estimated from 999 simulated data sets.

### 7.2.8 Application to CEDRS-like data

The data set used for this analysis is a randomly perturbed version of the CEDRS data for Colorado, with all personal identifiers removed. The diseases were identified by number. The analysis was performed for all of Colorado using (county centroid, month) coordinates as centroids. The 2000 census values for the county populations were used as the populations at risk, though this could be refined to reflect the etiology of particular diseases and updated population estimates. The study period was restricted to 1/1999 to 4/2004 to facilitate computation.

For disease code 17, both ClusterSeer and SaTScan produced the primary cluster of Boulder, Larimer, and Weld counties from 8/2001 through 3/2004, and a secondary cluster on the Eastern plains centered at the centroid of Cheyenne, including counties with centroids within approximately 200 kilometers of the Cheyenne centroid, and lasting from 3/2001 through 10/2003. A third cluster consisting of Adams county alone from 7/2003 through 9/2003 was also noted. These are in order of descending  $\lambda$  value. The secondary zone is restricted to zones that do not intersect the primary zone. Likewise, the third zone does not intersect the first two.

The p-values for all three zones were .001, as low as can be reported with 999 Monte Carlo simulations. Additional analyses are appended.

## 7.2.9 Recommendations

The Kulldorff spatio-temporal scan provides a sound method for locating clusters of distinctly elevated risk. The prospective version is well suited to the detection of localized outbreaks of a disease. Because the test identifies the most likely clusters (according to the maximum likelihood ratio), it can be used as an exploratory tool, whether or not the p-value falls below some cutoff.

The technique offers flexibility in the specification of expected regional risk, through the specification of the  $\mu(A)$ 's. The choice of zones can also be tailored to the geography and demographics of a region and the etiology of a disease.

The structure of the test minimizes concerns of multiple testing.

Some cautions when using this test should be noted. The technique can be computer-intensive for large data sets. Coarse aggregation in space or time to speed computation can obscure fine-scale clustering. Due to the broad alternative hypothesis, the power of the test to detect deviations from the null hypothesis is not available in closed form.

## 7.3 Knox Test

### 7.3.1 Introduction

Knox tests for space- time interaction are among the earliest clustering tests. In his 1964 paper "The Detection of Space Time Interactions" [?], E. G. Knox proposed a simple test to determine when the distribution of cases of a disease exhibits unusual clustering in space and time. Over the years statisticians have evaluated Knox's test and proposed variations and extensions of it. Since Knox's test is simple, reliable and widely used, it merits our serious consideration.

Knox's test has several variations. The simple tests use existing data to determine the existence of space-time interactions in the data. A space-time interaction occurs when points are closer together in time than one would expect if the times assigned to the data points were randomly permuted. It does not necessarily detect clusters of points. If all of the points in the data set are close together in space and time, the distribution of times might be about what you would expect regardless of how the times are permuted. If you want to detect clusters in space,

time, or space and time, use tests such as Kulldorff’s test for clusters in space and time or tests for clustering in space or clustering in time. See chapters 5 and 6 of this report.

A local Knox test assesses the existence of a space-time interaction at a particular point. A “prospective” version of Knox’s test uses new data along with previous data to determine whether space-time interactions are emerging over time. Both versions of Knox’s test use two time categories, near and far, to determine whether points are close together or far apart, but other versions of Knox’s tests can be designed to use multiple time intervals. We will consider each of these types of Knox’s test below.

### Notation

We will use the following notation to describe the tests:

$P_i : (x_i, y_i, z_i)$  A record of the  $i^{th}$  occurrence of a disease (the  $i^{th}$  case).

$s_{CRIT}$  : The critical distance in kilometers (a test parameter).

$t_{CRIT}$  : The critical time in days (a test parameter).

$n_s$  : The number of pairs of cases that are near one another in time (i.e., the two cases are within  $t_{CRIT}$  of one another).

$n_t$  : The number of pairs of cases that are near one another in space (i.e., the great arc length distance between the two cases is less than or equal to  $s_{CRIT}$ ).

$n_{s,t}$  : The number of *pairs* of cases that are near one another in both space and time.

$N_{s,t}$  : The random variable whose observed values are  $n_{s,t}$ .

$n$  : The total number of cases.

### 7.3.2 The $\chi^2$ Test

This test is simple. When used properly it is reliable. This test does not work well for rare diseases. If the number of cases in any cell is very small compared to the total number of pairs, it may report a space-time interaction when no space-time interaction has occurred. Use it only if the expected number of cases that are close in space and time is five or more, i.e.,  $E(n_{s,t}) \geq 5$ , where  $E(n_{s,t}) = \frac{n_s n_t}{n}$ , and when the number of cases in each cell is not very small compared to the total

number of cases. If your results from the chi square test do not agree with the results from the Poisson test, the Poisson test is likely to be more reliable. See Section 6 for more details. Given a set of  $n$  cases of a disease, the  $\chi^2$  test counts pairs  $(P_i, P_j)$  of cases that are close to one another in space, ( $distance(P_i, P_j) \leq s_{CRIT}$ ), close together in time, ( $|t_i - t_j| \leq t_{CRIT}$ ), or close to one another in both space and time. The counts are recorded in a contingency table.

Table 7.1: Observed Counts

<i>Time</i>	<i>Dist</i> $\leq$ $s_{CRIT}$	<i>Dist</i> $>$ $s_{CRIT}$	<i>Total</i>
<i>Time</i> $\leq$ $t_{CRIT}$	$n_{s,t}$	$n_t - n_{s,t}$	$n_t$
<i>Time</i> $>$ $t_{CRIT}$	$n_s - n_{s,t}$	$n - n_s - n_t + n_{s,t}$	$n - n_t$
<i>Total</i>	$n_s$	$n - n_s$	$n$

Table 7.2: Expected Counts

<i>Time</i>	<i>Dist</i> $\leq$ $s_{CRIT}$	<i>Dist</i> $>$ $s_{CRIT}$	<i>Total</i>
<i>Time</i> $\leq$ $t_{CRIT}$	$\frac{n_s n_t}{n}$	$\frac{(n - n_s) n_t}{n}$	$n_t$
<i>Time</i> $>$ $t_{CRIT}$	$\frac{(n - n_t) n_s}{n}$	$\frac{(n - n_s)(n - n_t)}{n}$	$n - n_t$
<i>Total</i>	$n_s$	$n - n_s$	$n$

$H_0$  : The time intervals between cases are independent of the space intervals between cases.

$H_1$  : The time intervals and the space intervals are not independent, (i.e., there is an interaction between space and time).

**Test Statistic** :  $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$ .

**Significance** : If  $\chi^2 > \chi_1^2(\alpha)$ , reject  $H_0$ .

For example, for  $\alpha = 0.05$ , the critical value of the  $\chi^2$  statistic is 3.88. Reject  $H_0$  if  $\chi^2 > 3.88$ , indicating a significant space-time interaction at the 5% level of significance.

### 7.3.3 The Poisson Test

Knox (1964) suggested that if  $N_{s,t}$  is small compared to  $n$ , it follows a Poisson distribution with expected value  $E(N_{s,t})$ . He formulated the following test based on the Poisson distribution:

$H_0$  : The time intervals between cases are independent of the space intervals between cases.

$H_1$  : The time and space intervals are not independent.

**Test Statistic** :  $N_{s,t} = n_{s,t}$ .

**Significance** : If  $P(X > N_{s,t}) < \alpha$ , reject  $H_0$ .

For the Poisson distribution, the expected value and the variance are equal. One way to check whether  $N_{s,t}$  follows a Poisson distribution is to calculate the variance independently and compare it to the expected value  $E(N_{s,t})$  given above. If the two numbers are close it may be reasonable to use the Poisson distribution; otherwise, the use of the Poisson distribution as the reference distribution may not be valid. The formula below, derived by Barton and David [?] has been cited by several authors (e.g., Rogerson 2001).

$$Var(N_{s,t}) = \frac{2n_s n_t}{n(n-1)} + \frac{4u_s u_t}{n(n-1)(n-2)} + \frac{4[n_s(n_s-1) - u_s][n_t(n_t-1) - u_t]}{n(n-1)(n-2)(n-3)} - \left(\frac{2n_s n_t}{n(n-1)}\right)^2 \quad (7.1)$$

where

$$u_s = \frac{1}{2} \sum_i n_s(i)^2 - n_s \quad (7.2)$$

and

$$u_t = \frac{1}{2} \sum_i n_t(i)^2 - n_t \quad (7.3)$$

### 7.3.4 Test using the Normal Distribution

If the number of cases is large and  $n_{s,t}$  is not too small, the normal approximation to the Poisson distribution can be used as the reference distribution for the Knox statistic under the null hypothesis (no clustering). This test has the following form:

$H_0$  : The time intervals between cases are independent of the space intervals between cases.

$H_1$  : The time intervals between cases are not independent of the space intervals.

**Test Statistic** :  $Z^* = \frac{n_{s,t} - E(N_{s,t})}{\sqrt{Var(N_{s,t})}}$ , where  $E(N_{s,t}) = n_s n_t / n$  (see Table 7.2) and  $Var(N_{s,t})$  is given in Eqn (7.1).

**Significance** : If  $P(Z > Z^*) < \alpha$  reject  $H_0$ , *e.g.*, if  $\alpha = 0.05$  and  $Z > 1.645$  reject  $H_0$ .

### 7.3.5 Monte Carlo Simulation

Uncertainty about the distribution of the Knox statistic has led some researchers to estimate its distribution by Monte Carlo simulation. One common procedure is to take a set of cases, fix the locations in space but permute the times, and then calculate  $N_{s,t}$  for each permutation. With a large number of cases, repeating the process for a large number of permutations gives a good approximation to the distribution of  $N_{s,t}$  for the given set of locations. ClusterSeer uses both the  $\chi^2$  distribution and the “random permutation” method (called “Monte Carlo Simulation” in the ClusterSeer manual), to assess the significance of the Knox test. Because ClusterSeer uses only 999 random permutations, the smallest significance level that can be observed is 0.001 (1 in 1000).

### 7.3.6 A Local Knox Test

P. A. Rogerson [?] formulated a Knox test for time and space interaction at a particular point. This version of the test is appropriate if interest lies in assessing the existence of a cluster at a particular point. To ensure statistical validity, the point must be chosen *before* looking at the data; *e.g.*, it can be chosen (otherwise, the significance level can be grossly understated if the data are used to decide on the particular point, or on repeated tests at multiple points). The test for space and time interaction at case  $P_i$  is given below.

Additional Notation:

$n_{s,t}(i)$  = number of cases that are close to case  $P_i$  in both time and space.

$n_s(i)$  = number of cases that are close to case  $P_i$  in space.

$n_t(i)$  = number of cases that are close to case  $P_i$  in time.

$N_{s,t}(i)$  = random variable whose values are  $n_{s,t}$ .

Rogerson determines the probability distribution of  $N_{s,t}$  using permutations. He considers a set of  $n$  cases with fixed positions and permutes the times  $(t_1, t_2, \dots, t_n)$ . He groups the permutations according to which time is assigned to case  $P_i$ , yielding  $(n-1)!$  permutations in each group. Let  $n_i^j$  be the number of cases close to case  $P_i$  when this case is assigned to time  $j$ . Under the null hypothesis, (given below), all permutations of time are equally likely. For the group of permutations that assign time  $j$  to case  $P_i$ , the distribution of  $N_{s,t}(i)$  is hypergeometric with parameters  $n-1, n_s(i), n_i^j(i)$ . This can be expressed as:

$P(N_{s,t}(i) = n_{s,t}^j) = Q_j$  , where

$$Q_j = \frac{\binom{(n_t)^j(i)}{n_{s,t}(i)} \binom{n-1-n_t^j(i)}{n_s(i)-n_{s,t}(i)}}{\binom{n-1}{n_s(i)}}$$

Assigning equal weights to each of the groups of permutations and summing gives the probability distribution of  $N_{s,t}$ .

$$P\{N_{s,t}(i) = n_{s,t}(i)\} = \frac{1}{n} \sum_{i=1}^n Q_j \quad (7.4)$$

Rogerson's Local Knox Test is as follows:

$H_0$  : There is no interaction between space and time at case  $P_i$ .

$H_1$  : There is an interaction between space and time at case  $P_i$ .

**Test Statistic** :  $N_{s,t}(i) = n_{s,t}(i)$

**Significance** :  $p$  - value =  $\sum_{k=n_{s,t}}^{\max(n_t^j(i))} P\{(N_{s,t}(i) = k)\}$

Rogerson also gives a normal approximation for this local test. The hypotheses are the same as they are for the local Knox test. The test statistic and the significance are given below:

**Test Statistic** :  $Z = \frac{[n_{s,t}(i) - E(N_{s,t}(i)) - 0.5]}{\sqrt{V(N_{s,t}(i))}}$

**Significance** : If  $Z > Z_{1-\alpha}$ , reject  $H_0$ .

Rogerson gives the following formulas for the expected value and the variance of the local test statistic  $N_{s,t}(i)$ .

$$E(N_{s,t}(i)) = \frac{2n_t n_s(i)}{n(n-1)}$$

$$V(N_{s,t}(i)) = 2(n-1)n_t - \frac{\sum_{j=1}^n (n_t^j(i))^2 n_s(i)(n-1-n_s(i))}{n(n-1)^2(n-2)}$$

### 7.3.7 A Prospective Knox Test

Rogerson (2001) presents a version of Knox's test that will allow the user to incorporate new data into the test as the new data come in and to monitor the changing values of the Knox statistic. A significant increase in Knox's statistic signals an emerging space time interaction. The hypotheses for this test are the same as the hypotheses for the local Knox test. The case in question is the new data point. The sample test statistic is  $K_i = K_{i-1} + n_{s,t}(i)$ . Rogerson uses the normal approximation to the  $K_i$  distribution to conduct the test. He calculates the standard  $Z$  statistic using the expected value and variance for  $K_i$  conditioned on the value of  $K_{i-1}$ ,  $n_s(i-1)$ , and  $n_t(i-1)$ . However he also proves that this  $Z$  value is equal to the  $Z$  value from the local Knox test. Thus we can follow the procedure for the local Knox test to assess the existence of an emerging cluster at case  $P_i$ .

Rogerson uses a cumulative sum  $S_j$  to monitor the  $K_i$  values,  $\sum_{i=1}^j k_i$ . The cumulative sum procedure has two user defined parameters,  $k$  and  $h$ . The parameter  $k$ , measured in standard deviations, determines how large the normalized sample test statistic must be to cause an increase in  $S_j$ . Some writers suggest taking  $k$  equal to one-half times the size of the change in  $K_i$  that one wants to detect. For example if you consider an increase in 1 standard deviation to be significant, set  $k = \frac{1}{2}$ . The parameter  $h$  indicates how large  $S_j$  must be before the test reports that a space-time interaction has occurred. Small values of  $h$  will tend to result in false alarms, while large values of  $h$  may result in missed interactions. Rogerson gives a table showing the average run length for different values of  $h$  under the null hypothesis. When there is no significant space-time interaction, using  $h = 2.5$  will result in a reported space-time interaction about once in every 69 observations. With  $h = 5.5$  the false-alarm rate drops to about once in every 1560 observations.

### 7.3.8 How to conduct a Prospective Knox Test

- (a) Choose the value of  $h$ . If the cumulative sum,  $S(j)$ , is greater than  $h$ , a significant increase in  $K_i$  has occurred.
- (b) Run the ordinary Knox test on the existing data.
- (c) Calculate the standard normal variate,  $Z = \frac{[n_{s,t} - E(N_{s,t}) - 0.5]}{\sqrt{Var(N_{s,t})}}$
- (d) Initialize  $S_0 = \max(0, Z - n_{s,t})$ .
- (e) For each new data point calculate  $n_{s,t}(i)$ .
- (f) For each new data point, calculate the new  $Z$  value,  $Z = \frac{[n_{s,t}(i) - E(N_{s,t}(i)) - 0.5]}{\sqrt{Var(N_{s,t}(i))}}$ .
- (g) Update  $S_i$  using the formula  $S_i = \max(0, S_{i-1} + Z_i - k)$ .
- (h) If  $S_i > h$  a significant increase in  $K_i$  has occurred. This may indicate the emergence of a new space-time interaction.

Rogerson (2001) applied this process to counts of cases of Burkitt's lymphoma in Uganda. The results generally agreed with the results of previous tests and it detected some emerging interactions.

### 7.3.9 Knox Tests with More than Two Time Intervals

Another generalization of Knox's test allows the user to specify the number of time intervals  $T_{MAX}$  as well as the number of space intervals  $S_{MAX}$ . Knox treated the resulting  $T_{MAX} \times S_{MAX}$  array as a contingency table and calculated a  $\chi^2$  statistic with  $(T_{MAX} - 1) \cdot (S_{MAX} - 1)$  degrees of freedom. Since the rows and columns of the array are cumulative the independence conditions required for a  $\chi^2$  test do not apply. O. Abe [?] proposed an alternate statistic that reportedly follows a  $\chi^2$  distribution. We were not able to get a copy of his dissertation in which this is proved. Abe's article does include an illustration that unfortunately contains some typographical errors in the calculations, but his theoretical calculation of the variance of Knox's generalized statistic is correct.

We do not recommend Knox tests that use more than one critical time value or critical space value. The entries in the cells cannot be independent as required for the null hypothesis because the entries are cumulative sums. The number of pairs that are closer in space than  $s_{crit_1}$  will also be counted among the pairs that are closer in space than

any larger  $s_{crit_2}$ . Also the test statistic will be sensitive to extreme values. One of the advantages of the original Knox test is that the test statistic is robust. Since there are only two categories, near and far, the test statistic  $n_{s,t}$  is not sensitive to extreme values.

### 7.3.10 Evaluation of Knox’s Test

We coded the three versions of the basic Knox test in the statistical programming language R and ran the code on data sets for diseases 15, 17, 28, and 87. We chose the parameters to be numbers we thought would be reasonable and would illustrate the behavior of the test under various conditions. We summarize the results below. To interpret this table, consider the entry for disease 15 with  $s_{crit} = 1km$ , and  $t_{crit} = 30days$  (second line): the chi-square test and the normal approximation test do not support a space-time interaction while the Poisson test does. We will discuss this discrepancy below.

Table 3: Results of Knox’s test using 3 possible reference distributions: chi-squared (§7.3.2), Poisson (§7.3.3), Normal (§7.3.4).

$(s_{crit}, t_{crit})$	Disease 15	Disease 17	Disease 28	Disease 87
(1, 7)	No, No, No	Yes, Yes, Yes	No, No, No	No, Yes, No
(1, 30)	No, Yes, No	No, No, No	No, No, No	No, No, No
(1, 90)	No, No, No	No, No, No	No, No, No	No, No, No
(10, 7)	No, No, No	Yes, Yes, Yes	No, No, No	No, Yes, No
(10, 30)	No, Yes, No	No, No, No	No, No, No	No, No, No
(10, 90)	No, No, No	No, No, No	No, No, No	No, No, No
(50, 7)	No, No, No	Yes, Yes, Yes	No, No, No	No, Yes, No
(50, 30)	No, Yes, No	No, No, No	No, No, No	No, No, No
(50, 90)	No, No, No	No, No, No	No, No, No	No, No, No

Overall, the three versions of Knox’s test give consistent results. For the data for disease 17 they indicate a possible space-time interaction for cases that occur within 7 days of each other. The Poisson test indicates likewise for disease 87 ( $t_{crit}=7$  days) and also for disease 15 when  $t_{crit}= 30$  days. No space-time interaction appears in the data for disease 28. In two instances the tests disagree. For disease 15 and  $s_{crit} = 30$ , the chi-square test and the Z test indicate no interaction

while the Poisson test shows a possible space-time interaction. In this case there are 3 pairs that are close in space and time. The p-values for the chi-square test and the Poisson test are 0.051 and 0.022 respectively and the p-value for the normal approximation is 0.07. While only the Poisson test is significant at the 5% significance level, the other two p-values are fairly close. All 3 p-values are rather small, which suggests evidence of space-time interaction. Likewise, with disease 87, the Poisson test indicates a possible interaction while the other two tests do not. In this case the total number of pairs that are close in space and time is 7. The P-values for the three test statistics are 0.10, 0.04 and 0.22 respectively. In this case a space-time interaction is less likely.

Although we are less certain of the result when we get conflicting information, we note that the Poisson test is generally more reliable than the  $\chi^2$  test when we have small numbers of cases in one of the cells of the contingency table. If we have reason to think that this is the case we can apply Fisher's Exact Test. The null hypothesis for Fisher's Exact Test is that the odds of a case occurring in any cell are the same, i.e., the odds ratio is equal to 1. This corresponds to the hypothesis that there is no space-time interaction. The procedure for conducting Fisher's Exact Test appears below. We ran Fisher's Exact test on the data from disease 15. The contingency table was the same for  $s_{crit} = 1, 10$  and 50 km. There were 3 pairs close in space and time. Fisher's test supported the hypothesis of no space-time interaction with Pvalue = 0.08576. We conclude that the apparent significance of the Poisson test statistic was due to the very small number of cases rather than to a significant space-time interaction. For disease 87 we also had the same contingency table for all three values of the distance parameter. In this case there were 7 pairs out of 3,760,653 that were close in space and time. The algorithm for Fisher's test would not accept such a large number of pairs. When we scaled the matrix to a size that the algorithm would accept, the number of pairs close in space and time scaled to 0. Although we were unable to run Fisher's test we conclude that there is likely no significant space-time interaction in disease 87.

To conduct Fisher's Exact Test proceed as follows:

- (a) Create a matrix containing the entries of the Chi Square contingency table. Use the functions `casecounts` and `paircounts` in the appendix to get  $n_s, n_t, n_{s,t}$  and  $n$ . If some of the entries are very large (of order  $10^6$  or so), scale them and round to

integer values.

- (b) Use the R command *fisher.test(matrix name)*.

The command will return a Pvalue, a statement of the alternate hypothesis, a 95% confidence interval for the odds ratio, and an estimate of the odds ratio. A small Pvalue supports space-time interaction.

### 7.3.11 Recommendations

We recommend the following tests for space-time interactions:

- (a) A standard Knox test using the chi-square distribution.
- (b) A standard Knox test using the Poisson distribution.
- (c) A standard Knox test using the Normal approximation to the Poisson distribution.
- (d) A local Knox test
- (e) A prospective Knox test.

We do not recommend the test with more than two time intervals for the reasons noted in section 5 above. The calculations are cumbersome. Also since we are testing for space-time interaction, we do not gain an appreciable advantage by using more than one critical time value and one critical distance value. The test with several time intervals also tends to be more sensitive to extreme data values than does the standard Knox's test. Cautions should be observed when using Knox's test. The standard Knox test is sensitive to the choices of the critical distance and the critical time parameters. Choosing values that are too large or too small will give unreliable results. These critical values should be chosen from knowledge of the disease or from past data. The test loses statistical validity if you use the current data set to determine the critical parameters. The process also loses validity if repeated tests are done on the same data using different critical parameters.

# Chapter 8

## Recommendations for Graphical Display of Results

### 8.1 Temporal Display

Although GIS may be useful for displaying the results of a statistical test for space-time clustering, the full range of its tools cannot be used to display the results of a statistical test for temporal clustering only. We propose here two alternative displays for representing the results of temporal tests such as those from CUSUM tests. The first is a bar or line graph, where the y-axis is the reciprocal of the test's p-value (so that highly significant time periods of clustering are indicated with high levels on the y-axis). Results of different statistical tests are easily shown on the same plot, using different colored lines and line types (solid, dot, dash, long-dash, etc.). An example using the county- and month-level data for disease XXX is shown in Figure 8.1. If numerous tests are used we recommend that a line graph be used to display the results where as if 3 to 4 tests are used a bar graph will work very nicely. These graph can be created using a spread sheet such as EXCEL. Shown is an example of how to represent temporal statistics using a line graph:

Figure 8.1: Temporal Line Graph

The "higher" the line is on the y-axis, the more significant the clustering is for the particular month in which it is highest. In Figure 8.1 "Test 1" has a high significance of clustering in month 6, conversely "Test 3" has highest clustering in month 7.

The second method is a graph that shows the number of cases at time  $t$ .

Figure 8.2: Temporal Results by Cases

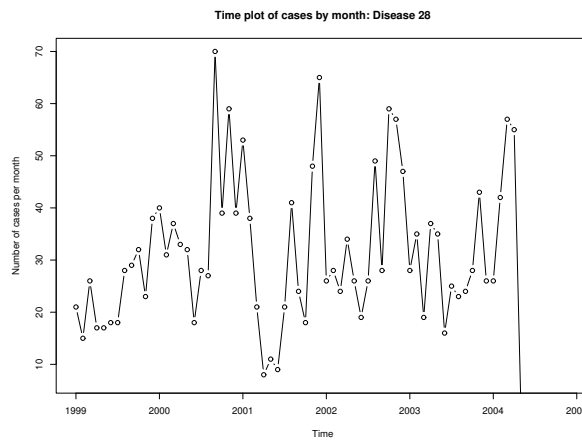
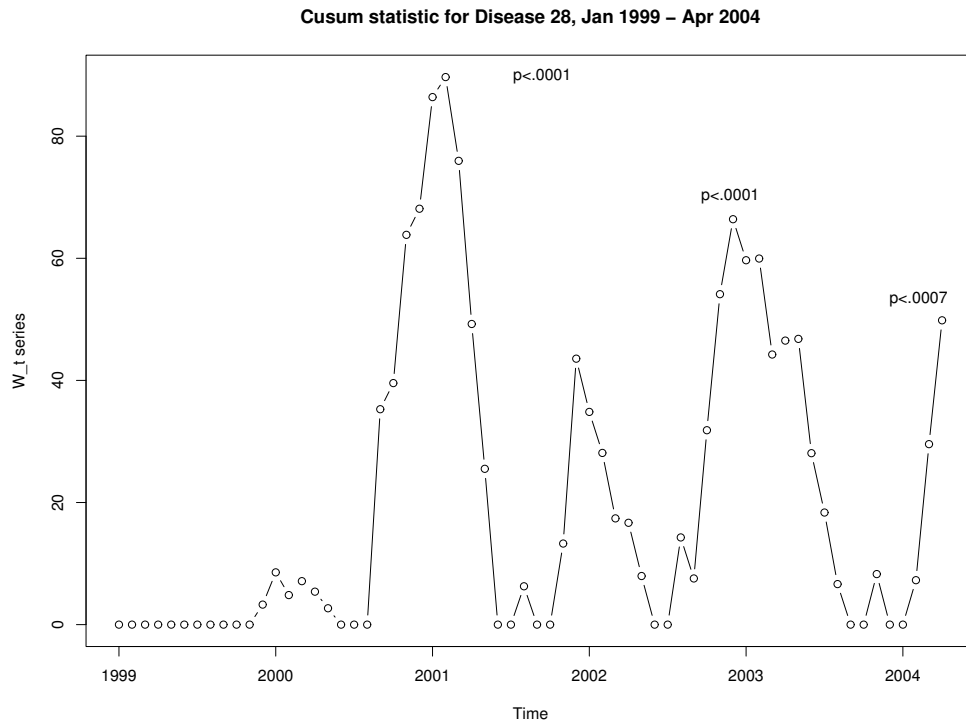


Figure 8.2 shows the method described above where by the y-axis represents the number of cases and the x-axis represents time. Also it is possible to use the y-axis to display the value of the CUSUM statistic and leave the x-axis the same.

Figure 8.3: Temporal Results by CuSum Statistics



The CUSUM statistic method is shown in Figure 8.3.

## 8.2 Spatial Display

We recommend the use of various geometric shapes located at the centroid of the outbreak coded in one color with varying shades proportional to the inverse of the significance. As an example, the smaller the significance level (i.e. little to no significance) the lighter the shade of the color. The ranges of the significance will be .10 to .06, .05, .04, .03, .02, .01, and less than .01. Figure 8.4 is a map of Colorado with spatial clustering:

Figure 8.4: Spatial Clustering

Figure 8.4 shows circles centered on the outbreaks, lists the significance of the outbreak, and shows the varying color scheme as discussed above. In the event that a circle is used to represent clustering, varying the radius as a function of the significance may be confusing and lead others to believe that the boundaries of the outbreak are that of the circle when in actuality the boundaries of the outbreak may in fact be much smaller or much larger for that matter. The reason for one color is the varying types of color blindness and the sensitivities to many colors (i.e. red, green, blue, and yellow). While the Atlas of United States Mortality [?] recommends against this, we feel that this is better suited for our purposes. The Atlas uses choropleth maps which separate regions based upon their similarity. This method does not seem as useful for the project since the results of the statistical tests give a level of statistical significance, rather than a region of significance.

For spatial analyses, a Geographical Information System (GIS), especially ARCGIS, is very useful for illustrating the regions where clustering is indicated. The results of the various statistical tests can be represented through the use of GIS layers. For example, Figure 8.5 shows the first layer:

Figure 8.5: Colorado Map With Counties

Figure 8.6 shows each individual case and could be the second layer:

Figure 8.6: Individual Cases

The next layers could represent each of the statistical tests and their calculated significance and region in Colorado:

Figure 8.7: Spatial Clustering

Consequently each layer may be superimposed on top of the Colorado map or may be viewed singularly with no map as shown below:

Figure 8.8: Spatial Clustering Layer 1

Figure 8.9: Spatial Clustering Layer 2

Figure 8.10: Spatial Clustering Layer 3

Figures 8.8, 8.9, and 8.10 display three layers that are constructed from the results of three different tests for clustering. Each test is illustrated with a different shape. Figure 8.8 displays the test<sub>1</sub> results using red circles; Figure 8.9 displays the test<sub>2</sub> results using blue hexagons; and Figure 8.10 displays the test<sub>3</sub> results using brown squares.”

### 8.3 Spatial-Temporal Display

For spatial and temporal representation, there is a practical way of using a GIS. The CDPHE uses a ”slide show” method currently to represent these results. They draw various maps in time and show the spatial clustering on those maps. ARCGIS can be suited for this particular method in question as well as the next method to be shown.

Figure 8.11: Spatial-Temporal Display

Figure 8.11 shows one method that can be used to show spatial-temporal statistics. Since the clustering occurs over a time span for a given set of counties, the method used to create Figure 8.11 is an example of a choropleth map (which observes the color scheme and rules noted above). It can provide a useful display of the results, because it indicates clustering for a given region. As with spatial statistics, this method can also be viewed in a layer by layer format.

Figure 8.12: Spatial-Temporal Display Layer

Figure 8.12 shows the actual layer not imposed upon the map, and thus this layer could be incorporated into the main map, along with all spatial results there by creating a map that has a maximum amount of information.

## 8.4 Conclusion

These recommendations are not exhaustive and merely are meant as an example for how such graphical models may be implemented. For further information on graphically displaying such results, using a GIS in particular, refer to the additional resources at [?] [?] [?] [?] in the bibliography.

# Chapter 9

## Summary of Recommendations

### 9.1 Comments on Software

#### 9.1.1 Cluster Seer

Though Cluster Seer has some advantages, we do not recommend it for the following reasons.

- Expensive
- No automation (i.e. each test has to have its own set up)
- Poor interface
- Data formats for different tests are not consistent
- County level data for Colorado was too large a data set and crashed the computer for one test.
- Unclear validity of the implementations of the tests, where distributional approximations are used (e.g. approximate critical points and p-values may not be uniformly valid for all the tests)

#### 9.1.2 Sat Scan

- Free (download from [www.satscan.org](http://www.satscan.org)).
- Relatively easy to run once the data are in the proper format

- Has fairly convenient formats for saving partial results, thus data file need not be totally reconstructed, if analysis is interrupted.
- Fast for county-month level data (< 1 minute)
- Easy to download and install
- Offers only scan statistic, but can test for clusters in time, space, time-space interaction.
- Includes facilities for including additional information beyond population at risk (e.g., expected number of torn knee ligaments could be based on both population at risk in certain age groups and on the number of ski resorts/hours).
- Ongoing research into further implementations (e.g., Sat Scan significance as determined from randomization tests is due to be published in 2005)

### 9.1.3 R

- Free (download from [www.cran.r-project.org](http://www.cran.r-project.org))
- Very flexible
- Easy to automate (routine daily or weekly runs)
- Requires programming skills in R
- Memory management is rather slow (especially for repetitive commands using “for” loops)

## 9.2 Comments on Methods

### 9.2.1 Temporal Clustering

- Appropriate for either diseases that exhibits no seasonality (e.g. E. Coli) or for diseases whose seasonal component has been either removed (e.g. example... ) or taken into account (e.g. CDC’s MMWR monthly ”Figure 1”)
- Standard control chart using CuSum or EWMA statistics, or Levin-Kline implementation of CuSum, are most sensitive for detecting gradual increase or trends in time.
- Standard Shewhart chart on number of cases (with limits of 23 time the square root of the average number of cases over time) are most sensitive for detecting sudden increases.

## 9.2.2 Spatial Clustering

- Cuzik-Edwards is preferable for global region(e.g. entire state) but not local (e.g. small area around, say, Rocky flats); (i.e. good for identifying existence of clustering, but not for identifying specific location of cluster)
- Besag-Newell needs to choose a risk for the population under consideration above which the test shows clustering. Choice of background risk is subjective and requires information about the specific population.
- Moran I and Geary Tests are tests good for identifying local clustering based on distances the user choose. The tests will become more powerful with knowledge of population.
- Kulldorff spatial scan statistic is both powerful and sensitive to disease clusters of various shapes.

## 9.2.3 Time and Space Interaction

- Kulldorff method is intuitive in its operation as well as flexible (shapes of clusters can be arbitrary, though circles are most common shape for most implementations, including Sat Scan program)
- Knox test can be recommended for testing for a space-time interaction, with the caveat that the result may depend on the probability distribution. The test assess whether the observed probabilities of clustering in time (i.e. the probabilities of cases being close/far in time/space) and in space are independent, or whether they are independent (i.e., tests for a space-time interaction, not for clustering per se).