LBL-29384

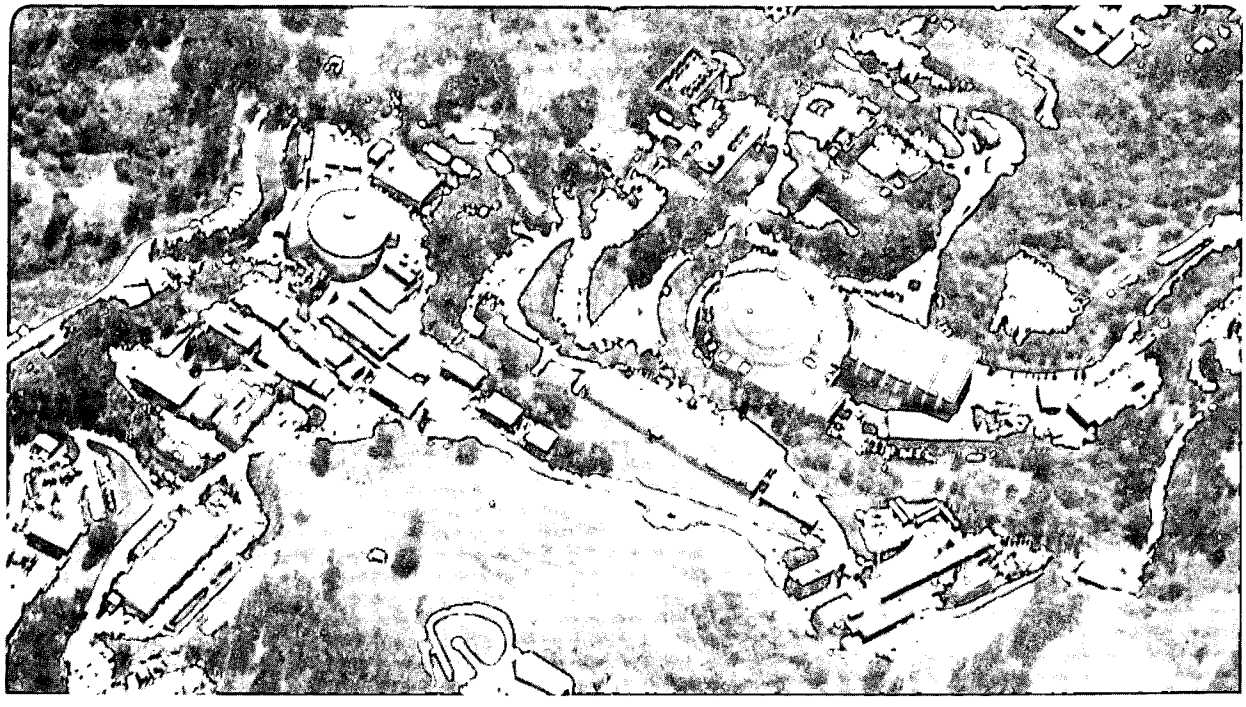
Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Information and Computing Sciences Division

The Spatial Evaluation of Neighborhood Clusters of Birth Defects

J.D. Frisch
(Ph.D. Thesis)

April 1990



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

DISCLAIMER

The Spatial Evaluation of Neighborhood Clusters of Birth Defects

Jonathan David Frisch
(Ph.D. Thesis)

Computing Sciences Research and Development
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

April 16, 1990

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, Human Health and Assessments Division of the United States Department of Energy under Contract No. DE-AC03-76SF00098.

The Spatial Evaluation of

Neighborhood Clusters of Birth Defects

The Spatial Evaluation of Neighborhood Clusters of Birth Defects

by

Jonathan David Frisch

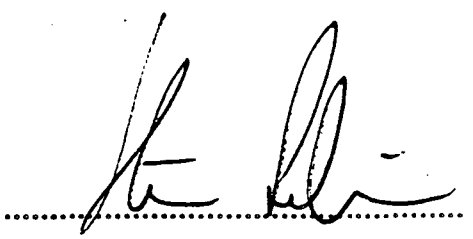
Abstract

Spatial statistics have recently been applied in epidemiology to evaluate clusters of cancer and birth defects. Their use requires a comparison population, drawn from the population at risk for disease, that may not always be readily available. In this dissertation the plausibility of using data on all birth defects, available from birth defects registries, as a surrogate for the spatial distribution of all live births in the analysis of clusters is assessed.

Three spatial statistics that have been applied in epidemiologic investigations of clusters, nearest neighbor distance, average interpoint distance, and average distance to a fixed point, were evaluated by computer simulation for their properties in a unit square, and in a zip code region. Comparison of spatial distributions of live births and birth defects was performed by drawing samples of live births and birth defects from Santa Clara County, determining the street address at birth, geocoding this address and evaluating the resultant maps using various statistical techniques. The proposed method was then demonstrated on a previously confirmed cluster of oral cleft cases. All live births for the neighborhood were geocoded, as were all birth defects. Evaluation of this cluster using the nearest neighbor and average interpoint distance statistics was performed using randomization techniques with both the live births population and the birth defect population as

comparison groups.

It was shown that comparing the nearest neighbor, average interpoint distance and average distance to a fixed point statistics from data thought to be clustered to the corresponding statistic generated from non-clustered data would detect clustering if a sufficient number of points were in the cluster, and provided the two samples were the same size. None of the statistics used detected a significant difference between the spatial distribution of the samples of all birth defects and of live births, although it was also demonstrated that the sensitivity of these methods to detect a systematic difference between the two populations was low. Reanalysis of the cluster, however, demonstrated that randomization analysis using all birth defects as a comparison population, was successful in detecting clustering.



Chair

Dedication

This dissertation is dedicated to the parents of children with birth defects, who face, on a personal level, the problems epidemiologists must consider in unemotional, clinical terms. The challenges they face must never be forgotten, and their quest for answers to the questions about birth defects must be supported and encouraged.

Acknowledgements

This dissertation was supported in part by the California Birth Defects Monitoring Program. I am grateful to Dr. John Harris and the other members of this program for their help and support. Additional resources were provided by the University of California and the Population at Risk to Environmental Pollution (PAREP) project of the Lawrence Berkeley Laboratory.

I am particularly indebted to Dr. Gary Shaw, who has influenced both my thinking about the epidemiology of clusters and my use of the English language, and Dr. Jane Schulman, who provided much-needed guidance into the use of spatial statistics. I am grateful to Professor Steve Selvin, who was patient with me when he didn't have to be, and who guided me unerringly through the process of creating a thesis. My academic advisor, Dr. Allan Smith insured that I remained an epidemiologist when the statistics became overwhelming, and provided much-needed moral support.

Dr. Eugene Hammel of the Demography Group at the University of California lent a different perspective to the problem of interpreting a cluster. Deane Merrill of the Lawrence Berkeley Laboratory provided some useful perspectives on mapping. Chapters 3 and 4 of this dissertation could not have been completed without the assistance of Ms. Louise Detweiler of the Vital Statistics Branch of Santa Clara County.

I thank my parents, Joseph and Joan Frisch, whose love and faith have kept me encouraged and focussed. Lastly, I thank my friends, particularly Lars Rohrbach, Bertha Sotello, Connie Long, Tom McDonald, and my many friends at the American Red Cross, for remaining friends.

Table of Contents

1. Introduction and Literature Review

1.1. Background

The investigation of clusters of disease has been one of the hallmarks of epidemiology since early times. A "cluster" of disease can be thought of as "a closely grouped series of events or cases of a disease...in relation to time or place, or both" [1]. The thought that some disease may result from the location where people live traces back to the time of Hippocrates, who writes of areas unfit to inhabit in "On Airs, Waters and Places" [2]. The first well known example of an investigation of a cluster was John Snow's work on a cholera outbreak in London in the 1800's [3]. As epidemiology is the study of the distribution of disease in populations, it is understandable how investigations of regions in which unusually large numbers of cases of disease are found would be of interest. Indeed, the term "epidemic," as commonly understood, refers to a clustering, in space and time, of some disease.

In this dissertation, methods of detecting or confirming the existence of spatial clustering will be examined for their suitability for use in investigating clusters of birth defects. Often, data on the spatial distribution of the underlying population at risk are difficult to obtain, particularly when the spatial data required involve more specific information on location than the census tract or county of residence. As clusters often occur in areas considerably smaller than such geopolitical areas, and as one cannot necessarily assume that the underlying population is uniformly distributed within these areas, it may be desirable to identify some surrogate population for which more precise data are available. A surrogate population may also be necessary in situations where no data exist on the locations of individuals in the

underlying population at risk, for example, when census tract or street address are not recorded. With the advent of population-based disease registries, such surrogate populations may be readily available within the registry responsible for investigating such clusters. As an example, the California Birth Defects Monitoring Program (CBDMP), a population-based disease registry that collects data on all structural birth defects manifest in the first year of life, will be used to illustrate the application of some of these spatial statistics.

In this chapter, spatial statistics that have been developed to investigate spatial clusters will be described, and some of the problems related to cluster investigations will be presented. Chapter 2 presents a series of simulations performed on some of these spatial statistics in the theoretical frame of a "unit square", and in a zip code region. Chapter 3 addresses the comparability of spatial location of birth defects to that of all live births in the population. The goal of this chapter is to ascertain whether the group of all birth defects could be used as a surrogate population for the true underlying population at risk, live births, in the ascertainment of spatial clustering of some specific birth defect. Chapter 4 will apply the spatial statistics and surrogate population to a previously investigated cluster of cleft lip and cleft palate that occurred between 1983 and 1986. In chapter 5, the applicability and generalizability of the results of the methods discussed will be presented.

A distinction is made between clusters occurring in space, those occurring in time, and those with both spatial and temporal association. This distinction, first made by Knox [4, 5], is important in both the consideration of the etiology of clusters, and in techniques used to ascertain them. Spatial clusters consider only the distribution of cases in a particular defined region,

with the only consideration of time being a definition by the investigator of a general period of interest.

Etiologically, diseases in which the temporal period is too long or short to be considered important would fit a spatial clustering model. A hypothetical example might be a cluster of birth defects caused by drinking water contaminated with some organic chemical in a single water district. With the comparatively small population at risk of pregnant women, and with a potentially extended period in which the drinking water may be contaminated, a close temporal association would not be expected to occur, and thus space-time methods would not detect a cluster. Spatial methods, however, might show a clustering of cases in that area served with the contaminated water.

Space-time clustering considers the interaction of both space and time. Proof of space-time aggregation of cases is often extremely revealing, as the disease rate is shown to vary in both time and place, which limits the possible explanations to a few variables that vary with the same pattern [6]. Etiologically, one might expect infectious agents, intermittent environmental exposure, or the introduction of a new drug with teratogenic properties and local popularity among prescribing physicians to produce evidence of space-time clusters.

Geographers have observed that there are two basic point pattern analysis methods that have been used extensively—nearest neighbor and quadrat methods [7]. Most of the space-time methods and spatial methods that calculate rates in adjacent spatial areas are actually forms of quadrat analysis. The average interpoint distance and distance to a fixed point methods are modified versions of nearest neighbor techniques. Neither of these methods compare the relative spatial locations of points, and as will

be seen in chapter 3, this may be difficult. Fortunately, the specific "shape" of the spatial distribution is usually not important to epidemiologists, who are more concerned with whether clustering is occurring.

In both spatial and space-time clustering, the smaller the geographic area of concern the stronger the potential for determining a cause for the outbreak, as fewer possible factors can be implicated. In addition, with space-time clustering, the shorter the time period of concern, the more powerful the analysis. Thus these methods are highly suited to detecting localized outbreaks, but become less powerful when dealing with larger more widely diffused increases in disease occurrence.

1.2. Cluster Investigations

Typically, cluster investigations are instigated as the result of reports made to a health department or disease registry from a concerned physician or a member of the community. Allegations of possible exposures thought to be the cause of the outbreak are often made with the initial cluster report.

The registry or other agency responding to the alleged cluster will first try to verify the existence of the "index cases", that is, those cases that are pointed to by the reporting party as evidence of an excess of disease. If the index cases are verified, additional case ascertainment is usually done, and then disease rates are calculated. Data on the population at risk, necessary for the calculation of rates, is usually derived from vital statistics sources. If the excess is significant, a case-control study may be initiated.

If exposure allegations are not made, the characterization of the area of concern may be subjective, and subject to a bias sometimes referred to as "Texas Sharp-shooting" [8], where the "bulls-eye" is drawn around the bullet-hole in the side of the barn. "Texas Sharp-shooting" can actually be

thought of as a form of exposure suspicion bias, since one is searching for exposure on the basis of a known outcome. In cluster investigations, this bias appears when the study area is created to be around the index cases, thereby increasing the probability of detecting a cluster in the analysis. To avoid this, and to permit the calculation of prevalence proportions using existing population data, investigators will often use a geopolitically defined region such as a census tract, zip code region, county, or other area. The drawback to using these areas is that they often contain a variety of population densities, and may contain sub-populations with different levels of risk for the disease of concern. While these are not problems when risk ratios are being calculated, they may cause a misinterpretation of spatial statistics such as the nearest neighbor distance.

Once additional case ascertainment is completed, the investigator must identify the population at risk for disease, and should ascertain the distribution of that population in the area of concern. Determination of the expected rate of disease in the population at risk and calculation of a standardized morbidity ratio will describe whether an excess is occurring. Comparison of the spatial arrangement of the disease cases and that of the underlying population at risk permits a statement of whether clustering of the disease of concern is occurring. This comparison of the spatial arrangement of disease is often not done in epidemiologic studies, because it requires specialized techniques that are not widely known. A list of techniques that have been applied in epidemiologic analysis appears in table 1.1. These methods have been described in detail elsewhere [9-11].

Table 1.1
Statistical Methods for Cluster Investigations

Method	Reference	
	Spatial Methods	
1.	Lloyd and Roberts	[12]
2.	Spatial Autocorrelation	[13]
3.	Grimson, Wang, and Johnson	[14]
4.	Ohno, Aoki and Aoki	[15]
5.	Nearest Neighbor Analysis	[16]
6.	Density Equalized Map Projections	[17]
	Space-Time Methods	
7.	Cell Occupancy Approach	[18]
8.	Pinkel, Dowd and Bross	[19]
9.	Ederer, Meyers and Mantel	[20]
10.	Knox	[21]
11.	Persisting Rates Approach	[22]
12.	Barton, David and Merrington	[23]
13.	Mantel	[24]
14.	Pike and Smith	[25]
15.	Smith and Pike	[26, 27]
16.	Goldstein and Cuzick	[28]
17.	Chen, *et al.*	[29]
	Temporal Methods	
18.	Bailar, Eisenberg and Mantel	[30]
19.	Scan	[31]
20.	Tango	[32]

1.3. Spatial Methods

Of the spatial methods listed in table 1.1, only methods 5 and 6, the nearest neighbor analysis introduced by Clark and Evans in the ecological literature in the 1950's, and the Density Equalized Map Projections of Selvin *et al*, introduced in the 1980's, make use of continuous measures without resorting to some form of dichotomization or categorization of distances into "close" and "not close" groupings.

1.3.1. Nearest Neighbor, r

Nearest neighbor analysis was adapted from the field of plant ecology, where it was first reported by Clark and Evans [16], and has recently been applied to epidemiologic analysis [33-35]. In this method, for each point in a set of n points, the nearest point is determined among the n-1 other points and these distances to the nearest neighbor are averaged to produce a measure of clustering ($\bar{r}$). For a given number of points n, the smaller the mean nearest neighbor distance is, the more likely clustering is going on. This mean distance has several useful properties. The variance of r can be estimated by use of equation 1.1. In addition, an expected value for the mean nearest neighbor may be calculated according to equation 1.2, where A is the area of the study region and n the number of points. This expected value requires the assumption of uniform distribution of the points in the study area. The variance of r, assuming a uniform distribution in space of the points, is given by equation 1.3.

$$\sigma^2_r = \frac{\sum(r_{obs}-\bar{r})^2}{n(n-1)} \tag{1.1}$$

$$E(\bar{r}) = \frac{1}{2}\sqrt{\frac{A}{n}} \tag{1.2}$$

$$Var(\bar{r}) = 0.068\frac{A}{n^2} \tag{1.3}$$

By calculating the ratio of observed to expected values of r, one may determine whether clustering is being exhibited. When this ratio is 1.0, it suggests that the data are spatially uniformly distributed, while a ratio less than one suggests clustering. This ratio can be tested by calculating g according

to equation 1.4 and comparing it to the gamma distribution for samples less than n=100.

$$g = \frac{\left[\frac{T_{observed}}{T_{expected}} - 1\right]\sqrt{n}}{0.52} \qquad (1.4)$$

The nearest neighbor method has several attractive properties: its lack of investigator-defined "closeness", its ease of application, and its ability to be compared to the normal distribution, if the assumption of uniform distribution of the underlying population at risk can be made. Disadvantages of the method include its partial dependence on the size of the study area, which affects the ability of the method to detect clusters if the boundaries have no biological basis or are selected randomly [36]. In addition, geographers have noted that the nearest neighbor metric does not reflect the exact pattern of points, but only their relative distances [37, 38].

Another problem with this statistic is that the values of the nearest neighbor distances to each point are not completely independent of each other. To understand why this is so, consider the extreme example shown in figure 1.1. In this example, the six points are arranged in pairs, such that the nearest neighbor for either point in each pair is the other point in the pair. Clearly the nearest neighbors of each point is dependent on the spatial location of the other point which will be serving as the nearest neighbor, and thus the assumption of independence required for applying a normal distribution to the nearest neighbor distance is strictly incorrect.

One drawback with this method is the need to establish boundaries for the points. This boundary can affect the expected value in some cases, as has been discussed by Clark and Evans [16] and by Shaw [11]. In addition, there appears to be a boundary effect on the data. This effect can be most

easily thought of by considering an area with several points in it, representing cases of a disease. An applied example might be a postal zip code region in a city, with points representing cases of some birth defect. If some of those points are near the boundary of the zip code region, their true nearest neighbor, that is, the case closest to them, could be outside the boundary of the study area, in some adjacent region not studied. This situation is analogous in many ways to a truncated data set. Since cases outside the region of concern are not included in the calculation of the nearest neighbor statistic, the calculated nearest neighbor distance for points near the boundary would actually be larger than the true nearest neighbor distance, since the distance would be calculated to the nearest neighbor within the census tract under study. The result is an overestimation in the ratio of observed:expected nearest neighbor distances. One suggested method of reducing this "boundary effect" has been to "wrap-around" the area, whereby the left boundary of an area is imagined to be contiguous with the right boundary, and the top with the bottom. Clark and Evans suggest that distances to nearest neighbors lying outside the area should be included, but that these points should not be used as index points in the calculation [16].

In this analysis, no correction for the boundary effect was incorporated, since in situations where the nearest neighbor statistic may be used for investigations of clusters, it is possible no data on nearest neighbors outside the study area are available. The use of a "wrap-around" method is not desirable, since it assumes the area adjacent to the study region has the same population density as that within the study area, which may not be the case.

1.3.2. Average Interpoint Distance, $\bar{i}$, $\bar{i}^2$, $\bar{i}_{harmonic}$

The average interpoint distance is a method derived from Knox's statistic [4], that was suggested by Lloyd and Roberts [12] [39] as a way of determining spatial proximity while controlling for the variation in population density of the population at risk. This method provides an estimate of the proximity of all points to each other, by calculating the $\frac{n(n-1)}{2}$ possible distances between points. The average squared interpoint distance is calculated by taking the square of each of the $\frac{n(n-1)}{2}$ distances, and then calculating the average. Lastly, the harmonic average interpoint distance is calculated by taking a harmonic mean of the average interpoint distances. The variances of $\bar{i}$ and $\bar{i}^2$ are calculated in the normal manner. The variance of $\bar{i}_{harmonic}$ may be estimated according to equation 1.5.

$$var(i_{harmonic}) = \frac{(i_{harmonic})^4}{i^4} var(i) \tag{1.5}$$

The expected value of $\bar{i}^2$ and its variance for the unit square when the points are uniformly distributed were calculated by Schulman [40] and are given by equations 1.6, and 1.7. The expected value of i, E(i), is approximately equal to $\sqrt{E(i^2)}$. The variance of $\bar{i}$ is approximated by equation 1.8 [40]. The values of $\bar{i}$ and $\bar{i}^2$ should be less than the expected value when clustering is present, as the clustered points will be closer together.

$$E(i^2) = 2\Big[V(X) + V(Y)\Big] \tag{1.6}$$

where: $V(X) = E(X^2)-(E(X))^2$, and in a unit square, $E(X)=0.5$.

$$Var(i^2) = \frac{1}{\binom{n}{2}}\Big[2(n-1)\Big[E(X^4)+E(Y^4)\Big]+4(n-1)\Big[E(X^2Y^2)-E(X^2)E(Y^2)\Big]+ \qquad (1.7)$$

$$8(n-1)\Big[E(X^2)\Big[E(Y)\Big]^2-E(X)E(X^3)\Big]-$$

$$8(n-1)\Big[E(Y)E(Y^3)+E(Y)E(X^2Y)+E(X)E(Y^2X)\Big]-$$

$$2(n-3)\Big[\Big[E(X^2)\Big]^2+\Big[E(Y^2)\Big]^2\Big]+8E(XY)\Big[2(n-2)E(X)E(Y)+E(XY)\Big]$$

$$-4(2n-3)\Big[\Big[\Big\{\Big[E(X)\Big]^2\Big\}^2-2\Big\{E(X^2)\Big[E(X)\Big]^2+E(Y^2)\Big[E(Y)\Big]^2\Big\}\Big]\Big]$$

$$Var(i) = \frac{Var(i^2)}{4E(i)^2} \qquad (1.8)$$

One property of simple mean distances is that all values, large and small, contribute the same, and indeed one or two extremely large values will skew the mean value. Mantel [24] suggested that by taking the reciprocal of distances, the short distances, characteristic of clustering, would be "spread out", while the range of great distances would be collapsed. This is desirable when a small number of cases are clustered, relative to the total number of cases, since it prevents the dilution of effect that otherwise takes place. As such, harmonic means were also calculated. The expected harmonic mean is unknown for i, but should be less than the observed mean.

1.3.3. Average Distance to a Fixed Point, $\bar{d}$, $\bar{d}^2$, $\bar{d}_{harmonic}$

The average distance to a fixed point ($\bar{d}$) has been suggested as a possible measure of clustering when the exposure allegation is that of a fixed point-source of pollution, such as a toxic waste site or smoke-stack [17, 40]. The mean $\bar{d}$ is calculated by taking the mean of the distances from each

point to the fixed source of exposure. The mean squared distance and harmonic mean are calculated in an analogous manner to the average inter-point distance statistics discussed above.

Expected values for $\bar{d}$ and $\bar{d}^2$ and their variances for the unit square, with no clustering and assuming a uniform population distribution, are known [40] and are calculated according to equations 1.9, 1.10, 1.11, and 1.12. In these equations $E(\bar{d})$ and $E(\bar{d}^2)$ are the expected values of the $\bar{d}$ and $\bar{d}^2$ statistics, and $E(X^n)$ and $E(Y^n)$ are the n^{th} moments of X and Y, which, in a unit square, may be calculated as $\frac{1}{2^{n+1}}$. Similarly, $E(X^n Y^m)$ may be calculated as $\frac{1}{n+1}\cdot\frac{1}{m+1}$ in a unit square. $x_o = y_o = 0.5$ in the unit square.

$$E(\bar{d}) = \sqrt{\mu}\left[1 - \frac{\sigma^2}{8\mu^2} - \frac{15}{128}\frac{\sigma^4}{\mu^4}\right] \tag{1.9}$$

$$\text{where } \sigma^2 = n\ Var(\bar{d}^2) \text{ and } \mu = E(\bar{d}^2)$$

$$Var(\bar{d}) = \frac{1}{n}\left[\frac{\sigma^2}{4\mu} + \frac{7}{32}\frac{\sigma^4}{\mu^3} - \frac{15}{512}\frac{\sigma^6}{\mu^5} - \frac{225}{16384}\frac{\sigma^8}{\mu^7}\right] \tag{1.10}$$

$$\text{where } \sigma^2 = n\ Var(\bar{d}^2) \text{ and } \mu = E(\bar{d}^2)$$

$$\bar{d}^2_{expected} = E(X^2) + E(Y^2) + x_o^2 + y_o^2 - 2[x_o E(X) + y_o E(Y)] \tag{1.11}$$

$$Var(\bar{d}^2) = \frac{1}{n}\Big[E(X^4) + E(Y^4) + 2E(X^2Y^2) - 4\Big[x_o E(X^3) + y_o E(Y^3)\Big] + \tag{1.12}$$

$$4x_o^2\Big[E(X^2) - (E(X))^2\Big] + 4y_o^2\Big[E(X^2) - (E(X))^2\Big] -$$

$$\Big[E(X^2) + E(Y^2)\Big]^2 + 4\Big[E(X^2) + E(Y^2)\Big]\Big[x_o E(X) + y_o E(Y)\Big] +$$

$$8x_o y_o\Big[E(XY) - (E(X)E(Y))\Big] - 4\Big[y_o E(X^2Y) + x_o E(Y^2X)\Big]\Big]$$

The pure temporal clustering methods listed in table 1.1, as well as other techniques that have been developed to deal with household or familial clustering [41-46], do not consider the spatial distribution of the cases in any way, and thus are not appropriate for the investigation of clusters occurring in neighborhoods or other spatially defined areas. As such, they will not be considered further here.

1.4. Cluster Investigations

The literature is replete with cluster investigations. Shaw [11] summarized several examples of clusters in the literature. Additional information on studies of space-time clustering may be found in Smith [10]. While most cluster investigations are performed without the use of spatial statistics, examples continue to appear in the literature of studies that make use of the Knox method [4], the nearest neighbor method [16], and even the method of Pike and Smith [25]. Recent examples of studies using these statistics include an investigation of nasopharyngeal carcinoma in Greenland Eskimos using Knox's method [47], an investigation of lymphoproliferative malignancies in young people using the nearest neighbor method [35], an investigation of cleft lip and palate in Oxford, England using Knox's method and the method of Smith and Pike [48], and others [49-55]. The major diseases studied in cluster investigations have continued to be cancer and birth defects, due in part to the lack of understanding of the causes of these disease, and to the hypotheses of various possible environmental risk factors, and in part to their observed clustering in communities.

The advent of disease registries for such outcomes as birth defects and cancer has also contributed to the large number of cluster investigations on these outcomes, by providing data and a standardized data collection

scheme that can be used for investigating and verifying or refuting the existence of a cluster quickly and efficiently, as opposed to the special studies that formed most of the cluster literature in the 1960's and 1970's. In addition, disease registries have made possible the utilization of systematically collected spatial data for both case definition, and for characterization of the reference population. In addition to providing accurate "baseline" data on disease rates in different communities, the spatial distribution of other, non-clustered cases in the registry could be readily used to assess whether clustering is occurring. Such use makes a fundamental assumption that the population of non-clustered disease cases in the registry is randomly distributed among all the underlying population at risk.

The CBDMP had performed over 61 cluster investigations between 1981 and June, 1987 in response to community reports. Of these, only 7 have had statistically significant relative risks ($p<0.05$), although studies were ongoing for 10 at the time of report [56]. Three of these clusters resulted in full epidemiologic study, of which none conclusively implicated an environmental cause. Many of these clusters are still being watched, however, as if an elevated rate continues to persist, further investigation may be warranted. Schulte [57] reported on 61 occupational clusters investigated by the National Institute for Occupational Safety and Health between 1978 and 1984. While occupational clusters differ from spatial clusters in many respects, it is interesting to note that a similar lack of significant findings has resulted. Among the 61 investigations, 8 (13%) detected a statistically significant excess of disease, of which none had a plausible occupational etiology. Warner and Aldrich, in a 1988 article, noted that cancer cluster investigations undertaken by state health departments have been generally unproductive in the identification of possible environmental causes of cancer

[58].

1.5. Use of Surrogate Populations in Epidemiology

Using birth defects as an example, one may use the population of all birth defects as a surrogate for the true population at risk for having a birth defect (the population of all live births) if one can assume that no environmental or other factor causes *all* birth defects to distribute themselves in space in some way other than live births do. If one accepts the postulate that there are no universal teratogens (that is, no chemical or other agents that cause *all* birth defects without regard to type), then this assumption may be accepted. If a teratogen were to change the spatial distribution of some subset of defects, including the one under investigation, it would reduce the ability to detect a cluster. A major objective of this dissertation will be to attempt to show that all birth defects can be used as a surrogate for all live births in a cluster investigation, and to demonstrate this method on a previously identified cluster.

There are many examples of the use of surrogate populations for the underlying population at risk in epidemiology. Many case control studies make use of "diseased controls" such as those in the hospital for some other reason than that being studied, other non-related cancer cases, or users of a particular clinic or physician. The use of these "diseased controls" has been discussed elsewhere [59,60]. The major concern voiced over using a surrogate comparison population is the issue of bias. Most commonly Berkson's bias is cited as a major problem, that is, that ill populations are fundamentally different from the general population in some way, and that their use as a referent group is therefore biased, with respect to exposure. The equivalent of Berkson's bias in a cluster investigation using a

surrogate population would be if those diseased controls distribute themselves in space differently than the true referent population in a manner affected by the exposure of interest. Since it is unlikely a single environmental factor may increase risk for *all* birth defects, it is probable that this bias will not be a significant problem in the use of all birth defects as a surrogate population. Other biases discussed related to the use of diseased controls [59], including problems with case ascertainment, recall and interviewer bias are not generally of concern for the initial investigation of clusters.

1.6. Bias in Cluster Investigations

Cluster investigations are generally started in response to the concerns of the public or of medical professionals that some environmental factor may be causing an excess of disease in some area. Using community reports of suspected clusters generates a bias of its own, as has been discussed by Rothman [6]. The problem inherent in using community-reported clusters is a strange one: because community members were *looking* for a cluster and found one introduces a bias known as "multiple-comparisons" or "data-dredging" bias.

The fundamentals of data-dredging bias are that if one goes out looking for a statistically significant relation at some level, if one looks at enough relations one or more of them will be statistically significant just by chance. As an example, if one considers 100 communities looking for an excess of some birth defect significant with a p-value of 0.05, five of these 100 communities will be expected to show a statistical excess just by chance. Data-dredging bias in cluster reports is more vague; members of the public can be considered to be constantly making comparisons of rates, looking for

diseases for which incidence or prevalence appears to be elevated, compared to the "usual" rate. If a "significant" elevation is observed or suspected, it will be reported. In this way, one can consider the public to be making thousands of comparisons, since any combination of two or more events that are proximal in space may be compared to baseline rates. This form of multiple comparisons is intricately related to the definition of a cluster, as a "cluster" of one case will not be reported, environmental hypotheses not withstanding. An illustration of how important this data-dredging can be appears in chapter four, where the number of expected clusters of the same magnitude as the one illustrated is estimated for the entire state of California during the same period.

It is possible to characterize the period during pregnancy that a fetus is susceptible to developing particular defects, and thus to determine when the mother would have to be exposed to the exposure of concern in order to develop the outcome of concern. For many defects, this period of susceptibility is during the first trimester, and thus for a particular cluster investigation it is possible to define fairly exactly when exposure must have occurred in order for a particular defect to result. However, Kipen and Wartenberg note that when exposure assessment is performed retrospectively, power and precision may be lost due to recall and other biases [61]. When this is the case, clusters may appear to have no environmental explanation, since no common exposure can be confirmed.

Another bias that is introduced in the investigation of spatial clusters of birth defects is migration among the mothers during pregnancy. As many teratogenic agents have their most deleterious effects during the first trimester of pregnancy, using residence at birth may be a poor measure of exposure during the critical time period, if the mother moved during her

pregnancy. Khoury et al [62], using data from the Maryland Birth Defects Registry, observed that 20% of mothers with infants born with one of 12 sentinel defects had moved during pregnancy. Over four percent of the mothers of children in the registry had moved to a different county from their residence at conception. It was not possible to determine from Khoury's paper whether this migration pattern was consistent with that of mothers of children born without a birth defect. If mothers living near toxic waste sites are more likely to move when the find they are pregnant than other mothers, and if living near a toxic waste site is a risk factor for some birth defects, it is conceivable that, since pregnancy is not confirmed until after the first month or so post-conception, a differential bias may result.

Despite the problems inherent in cluster investigations, important and meaningful exposure-response relationships have been identified through this process. From contaminated drinking water in London that Snow showed to be the agent for the cholera outbreak [63], to a recent investigation in which erionite, a mineral, was shown to have carcinogenic properties [64], cluster investigations have proven to be useful. Their additional, often undocumented, utility as agents for developing causal hypotheses that can later be tested in more controlled epidemiologic settings further strengthens the need to continue rational, scientifically rigorous investigations of clusters. This dissertation will provide an additional tool that can be used for such investigations, that will further assist researchers in understanding clusters they may investigate.

1.7. References

1. Last, J. M. Dictionary of Epidemiology Second Edition. New York: Oxford University Press; 1988.
2. Hippocrates: Loep Classical Library Edition. Cambridge, MA: Harvard University Press; 1923: 66-137.
3. Snow, J. Snow on Cholera. New York: The Commonwealth Fund; 1936.
4. Knox, G. Epidemiology of childhood leukaemia in Northumberland and Durham. *Brit J Prev Soc Med.* 18: 17-24; 1964.
5. Knox, E. G. Epidemics of rare diseases. *Br Med J.* 27: 43-47; 1971.
6. Rothman, K. J. Clustering of disease [editorial]. *Am J Public Health.* 77: 13-15; 1987.
7. Gesler, W. The uses of spatial analysis in medical geography: a review. *Soc Sci Med.* 23: 963-973; 1986.
8. Grufferman, S. Hodgkin's disease. In: Schottenfeld, D.; Fraumeni, J. F., eds. *Cancer Epidemiology and Prevention.* Philadelphia: WB Saunders Company; 1982: 744.
9. Williams, G. W. Time-space clustering of disease. In: Cornell, R. G., ed. *Statistical Methods for Cancer Studies.* New York: Marcel Dekker Inc; 1984: 167-227.
10. Smith, P. G. Spatial and temporal clustering. In: Schottenfeld, D.; Fraumeni, J. F., eds. *Cancer Epidemiology and Prevention.* Philadelphia: WB Saunders; 1982: 391-407.
11. Shaw, G. M. A comparison of techniques used for the detection of spatial and temporal-spatial disease clustering. Doctoral Thesis.

Berkeley, CA: University of California; 1986.

12. Lloyd, S.; Roberts, C. J. A test for space clustering and its application to congenital limb defects in Cardiff. *Br J Prev Soc Med.* 27: 188-191; 1973.

13. Glick, B. The spatial autocorrelation of cancer mortality. *Soc Sci Med.* 13D: 123-130; 1979.

14. Grimson, R. C.; Wang, K. C.; Johnson, P. W. C. Searching for hierarchical clusters of disease: spatial patterns of sudden infant death syndrome. *Soc Sci Med.* 15D: 287-293; 1981.

15. Ohno, Y.; Aoki, K.; Aoki, N. A test of significance for geographic clusters of disease. *Int J Epidemiol.* 8: 273-281; 1979.

16. Clark, R. J.; Evans, F. C. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology.* 35: 444-453; 1954.

17. Selvin, S.; Merrill, D.; Schulman, J.; Sacks, S.; Bedell, L.; Wong, L. Transformations of maps to investigate clusters of disease. *Soc Sci Med.* 26: 215-221; 1988.

18. Pinkel, D.; Nefzger, D. Some epidemiological features of childhood leukemia in Buffalo NY area. *Cancer.* 12: 351-358; 1959.

19. Pinkel, D.; Nefzger, D. Some epidemiological features of childhood leukemia in the Buffalo, N.Y. area. *Cancer.* 16: 28-33; 1963.

20. Ederer, F.; Myers, M.; Mantel, N. A statistical problem in space and time. Do leukemia cases come in clusters? *Biometrics.* 20: 626-638; 1964.

21. Knox, G. Detection of low intensity epidemicity: application to cleft lip and palate. *Br J Prev Soc Med.* 17: 121-127; 1963.

22. Mustacchi, P. Some intracity variations of leukemia incidence in San Francisco. *Cancer.* 18: 362-368; 1965.

23. Barton, D. E.; David, F. N.; Merrington, M. A criterion for testing contagion in time and space. *Ann Hum Genet.* 29: 97-102; 1965.

24. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220; 1967.

25. Pike, M. C.; Smith, P. G. Disease clustering: a generalization of Knox's approach to the detection of space-time interactions. *Biometrics.* 24: 541-556; 1968.

26. Smith, P. G.; Pike, M. C. Space-time clustering: a case control method of examining diseases with long latent periods. In: Doll, R.; Vodopija, I., eds. *Host, Environment Interactions in the Etiology of Cancer in Man.* Lyon: IARC; 1973: 35-38.

27. Pike, M. C.; Smith, P. G. A case-control approach to examine diseases for evidence of contagion, including diseases with long latent periods. *Biometrics.* 30: 263-279; 1974.

28. Goldstein, I. F.; Cuzick, J. Application of a time-space clustering methodology to the assessment of acute environmental effects of respiratory illnesses. *Rev Environ Health.* 3: 259-275; 1981.

29. Chen, R.; Mantel, N.; Connelly, R. R.; Isacson, P. A monitoring system for chronic diseases. *Meth Inform Med.* 21: 86-90; 1982.

30. Bailar, J. C.; Eisenberg, H.; Mantel, N. Time between pairs of leukemia cases. *Cancer.* 25: 1301-1303; 1970.

31. Wallenstein, S. A test for detection of clustering over time. *Am J Epidemiol.* 111: 367-372; 1980.

32. Tango, T. The detection of disease clustering in time. *Biometrics.* 40: 15-26; 1984.

33. Lewis, M. S. Nearest neighbor analysis of epidemiological and community variables. *Psych Bull.* 85: 1302-1308; 1978.

34. Lewis, M. S. Spatial clustering of childhood leukemia. *J Chronic Dis.* 33: 703-712; 1980.

35. Freeman, A. C.; Zucker, R. S. An investigation into a reported cancer cluster using the nearest-neighbor statistical technique. *Prog Clin Biol Res.* 216: 53-58; 1986.

36. Pinder, D. A.; Witherick, M. E. The principles, practice and pitfalls of nearest neighbor analysis. *Geography.* 57: 277-288; 1972.

37. Vincent, P. Methodology by example: caution towards nearest neighbors 1. The general case: how not to measure spatial point patterns. *Area.* 8: 161-163; 1976.

38. Dawson, A. H. Are geographers indulging in a landscape lottery? *Area.* 7: 42-45; 1975.

39. Smith, P. G.; Pike, M. C. A note on a "close pairs" test for spatial clustering. *Br J Prev Soc Med.* 28: 63-64; 1974.

40. Schulman, J. The Statistical Analysis of Density Equalized Map Projections Ph.D. Thesis. Berkeley, CA: University of California; 1986.

41. Pearson, K. Multiple cases of disease in the same house. *Biometrika.* 9: 28-33; 1913.

42. Mathen, K. K.; Chakraborty, P. N. A statistical study on multiple cases of disease in households. *Sankhya.* 10: 387-392; 1950.

43. Walter, S. D. On the detection of household aggregation of disease. *Biometrics.* 30: 525-538; 1974.

44. Smith, P. G.; Pike, M. C. Generalisations of two tests for the detection of household aggregation of disease. *Biometrics.* 32: 817-828; 1976.

45. Fraser, D. W. Clustering of disease in population units: an exact test and its asymptotic version. *Am J Epidemiol.* 118: 732-739; 1983.

46. Ross, R.; Dworsky, R.; Paganini-Hill, A. The occurrence of multiple lymphoreticular and hematological malignancies in the same household. *Br J Cancer.* 47: 853-856; 1983.

47. Albeck, H.; Coleman, M.; Nielsen, N. H.; Hansen, H. S.; Hansen, J. P. H. Space-time clustering of nasopharyngeal carcinoma in Greenland Eskimos. *Br J Cancer.* 52: 909-914; 1985.

48. Heath, A. B. Cleft lip and palate in the Oxford Area; an examination of the evidence for clustering in space and time. *Br J Prev Soc Med.* 31: 269-271; 1977.

49. Silcocks, P. B. S.; Murrells, T. Space-time clustering and bone tumours: application of Knox's method to data from a population-based cancer registry. *Int J Cancer.* 40: 769-771; 1987.

50. Pinder, D. C. Trends and clusters in leukaemia in Mersey region. *Community Med.* 7: 272-277; 1985.

51. Mangoud, A.; Hillier, V. F.; Leck, I.; Thomas, R. W. Space-time interaction in Hodgkin's disease in Greater Manchester. *J Epidemiol Community Health.* 39: 58-62; 1985.

52. Hills, S. E.; Parkes, S. A.; Baker, S. B. Epidemiology of sarcoidosis in the Isle of Man--2: evidence for space-time clustering. *Thorax.* 42: 427-430; 1987.

53. Siemiatycki, J.; Colle, E.; Aubert, D.; Campbell, S.; Belmonte, M. M. The distribution of type I (insulin-dependent) diabetes mellitus by age,

sex, secular trend, seasonality, time clusters, and space-time clusters: evidence from Montreal, 1971-1983. *Am J Epidemiol.* 124: 545-560; 1986.

54. Kyyronen, P.; Hemminki, K. Gastro-intestinal atresias in Finland in 1970-79, indicating time-place clustering. *J Epidemiol Comm Health.* 42: 257-265; 1988.

55. Day, R.; Ware, J. H.; Wartenberg, D.; Zelen, M. An investigation of a reported cancer cluster in Randolph, Massachusetts. *J Clin Epidemiol.* 42: 137-150; 1989.

56. California Birth Defects Monitoring Program. CBDMP Progress Report; July 1, 1985 - June 30, 1987. Berkeley, CA: State of California, Department of Health Services; 1987.

57. Schulte, P. A.; Ehrenberg, R. L.; Singal, M. Investigation of occupational cancer clusters: theory and practice. *Am J Public Health.* 77: 52-56; 1987.

58. Warner, S. C.; Aldrich, T. E. The status of cancer cluster investigations undertaken by state health departments. *Am J Public Health.* 78: 306-307; 1988.

59. Smith, A. H.; Pearce, N. E.; Callas, P. W. Cancer case-control studies with other cancers as controls. *Int J Epidemiol.* 17: 298-306; 1988.

60. Norell, S. E.; Ahlbom, A. Hospital versus population referents in two case-referent studies. *Scand J Work Environ Health.* 13: 62-66; 1987.

61. Kipen, H.; Wartenberg, D. Don't close the door: some observations on cancer cluster investigation. *J Occup Med.* 80: 661-662; 1988.

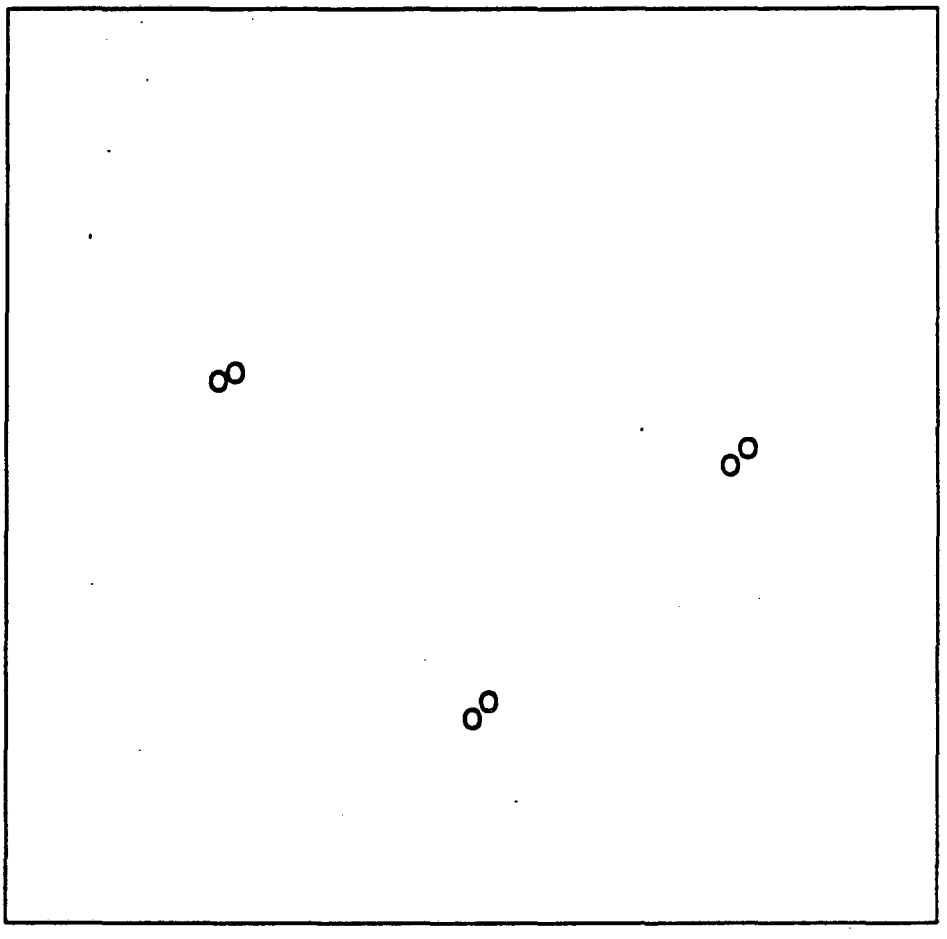
62. Khoury, M. J.; Stewart, W.; Weinstein, A.; Panny, S.; Lindsay, P.; Eisenberg, M. Residential mobility during pregnancy: implications for

environmental teratogenesis. *J Clin Epidemiol.* 41: 15-20; 1988.

63. Snow, J. On the mode of communication of cholera (2nd edition). New York: Haffner; 1965.

64. Baris, I.; Simonato, L.; Artvinli, M.; Pooley, F.; Saracci, R.; Skidmore, J.; Wagner, C. Epidemiological and environmental evidence of the health effects of exposure to erionite fibres: a four-year study in the Cappadocian region of Turkey. *Int J Cancer.* 39: 10-17; 1987.

Figure 1.1
Illustration of Non-Independence
of Nearest Neighbor Distances



2. Monte Carlo Simulations

2.1. Introduction

Evaluations of disease clusters have been accomplished using a variety of statistics. Many of the traditional methods, such as those described by Knox [1] and Mantel [2], evaluate space-time clustering, where cases must be close in both space and time for clustering to be detected. In addition to the possible problems of investigator bias inherent in the definition of clustering when using these statistics, space-time methods will not detect clusters that occur over an extended time period, or for which data collection is still in progress, as inter-event time periods cannot be completely characterized. When a disease has a long period between exposure and the first manifestations of illness (a long latency period), space-time methods may also fail to detect clustering. This is because if both the latency and the study period cover large periods of time, the clustering will appear to occur only in space. Also, as latency gets longer, the variability in time between exposure and disease onset can vary more broadly between individuals than in acute disease. This variability will further reduce the power of space-time methods to detect clustering, as it will reduce the apparent temporal clustering. In addition, in situations where an ongoing environmental insult is the putative cause of the outcome of concern, and where the outcome is considered endemic, the temporal aspect of clustering becomes irrelevant.

For such situations, statistics which consider only spatial variation are better suited for the evaluation of clusters. Three such statistics are the average nearest neighbor distance, the average distance to a fixed point, and the average interpoint distance.

To evaluate the performance of these spatial measures to detect clusters of birth defects, Monte Carlo simulations were performed under four spatial scenarios: a unit square with and without clustering, and a postal zip code region, with and without clustering. The first and second moments of the distribution of the statistics given were then compared.

For each simulated set of points, the statistics were calculated three ways: directly, in their usual ways, which will be referred to as the "observed" values, and by Jackknife and Bootstrap estimation procedures. As distance metrics, such as the nearest neighbor distances, between points are not completely independent, the mathematics of standard estimation techniques, such as maximum likelihood, are complicated. As such, the Jackknife and Bootstrap methods, which provide non-parametric estimates of the expectation and variance, were also employed.

2.2. Statistical Methods

The nearest neighbor distance, the average distance to a fixed point statistics, and the average interpoint distance statistics, as presented in chapter 1, will be used for these simulations.

2.2.1. Jackknife Technique

The Jackknife technique was introduced by Quenouille and Tukey in the 1950's [3-6], and is one of a series of methods for analyzing data for which the distribution is not known, or when the assumption of normality cannot be made. The Jackknife technique applied to a set of n points involves creating a series of n sub-populations, each with (n-1) points, by removal of a single point. So, for example, consider a set of three points $\{p_1,p_2,p_3\}$. To perform a

Jackknife analysis, three subpopulations would be generated: the first with p_1 removed, the second with p_2 removed, and the third with p_3 removed. For each of these three subpopulations, a summary statistic, such as the average interpoint distance ($\bar{i}$) is calculated, generating n estimates of the summary statistic that can then be recombined as an estimate of the actual statistic value.

A useful feature of the Jackknife method is that for means, the Jackknife mean will be equal to the observed mean. Thus, the Jackknife method perfectly estimates the observed value. The standard errors of these means will also be identical. This equality is *not* true for harmonic means, and thus the Jackknife method is retained as a possible estimator for the harmonic means of the statistics under consideration.

2.2.2. Bootstrap Technique

The Bootstrap method was first proposed by Efron in 1977 [7,8] and is a member of a family of nonparametric statistics that include the Jackknife, cross-validation, and half-sampling [8], all of which provide non-parametric estimates of standard error. The premise behind the Bootstrap is conceptually simple: a usually small sample of a larger population is available, and it is desired to estimate confidence intervals representing how well a statistic taken from the sample population reflects the true (unknown) population. While other methods are available for calculating this error, they make assumptions about the population sampled that may not be justified (e.g. normal distribution). To perform a Bootstrap analysis, the sample population (size n) is sampled repeatedly with replacement, generating "Bootstrap samples," each of size n, for which a summary statistic is then calculated. The resulting distribution of sample statistics (a "Bootstrap distribution") can be

used to estimate p-values. As the same n points from the sample population are used to generate the distribution, the distribution is said to be "pulled up by its own Bootstraps", hence the name.

Efron recommends sampling of up to 1000 Bootstrap samples in generation of the distribution, and while this job could not have been conceived of 30 years ago, it can now be accomplished readily with the use of a microcomputer [9].

A few examples exist of the use of the Bootstrap in epidemiology or related medical sciences. A recent paper presented at the 1988 annual APHA meeting in Boston, MA illustrated the Bootstrap in use to compare the reliability of laboratory tests used to monitor non-insulin-dependent diabetics [10]. Abbott and Carroll used the Bootstrap in a theoretical paper examining the interpretation of multiple logistic regression coefficients in prospective observational studies [11]. Weiss *et al* used the Bootstrap to determine the variability of regression coefficients in a study of the relationship between blood lead and blood pressure [12], and Mapleson has used the Bootstrap in assessing a clinical trial comparing Temazepam to Trimeprazine in the premedication of children for tonsillectomy [13]. Recently, the Bootstrap has been used to verify confidence intervals in a study of the associations of dietary fat, regional adiposity, and blood pressure [14].

The Bootstrap has been proposed for this research for several reasons. The lack of assumptions surrounding the nature of the statistics used (e.g. normal distribution) will allow the use of a wider variety of statistics, possibly including some that are not mathematically well defined. In addition, as the problem of small sample size is common in cluster investigations, assumptions that become unstable with small sample size will be less of a concern when the Bootstrap is employed.

2.3. Unit Square Analysis

The first series of simulations made use of a unit square, that is, a square of dimension 1 arbitrary unit in both the x and y direction. Uniformly distributed points were randomly placed into this square, and the statistics were calculated. The number of points used was varied and was selected to approximate typical numbers of cases found in the evaluation of clusters of birth defects. Four analyses were performed with 4, 6, 10, and 20 points, respectively.

The unit square can be thought of as a space in which the underlying population at risk is distributed uniformly, analogous to space in the "Density Equalized Map Projections" of Selvin, *et al.* [15]. While this condition is not expected in geographic areas such as zip code regions or census tracts, it permits an evaluation of the statistics without concern for where the underlying population reside, and therefore permits the calculation of expected values.

Expected values of the five statistics in a unit square were calculated according to equations 1.2, 1.5, 1.7, 1.9 and 1.11. For the moments, E(X), and E(Y) are taken to be 0.5, and the fixed point used in the average distance to a fixed point statistics, (x_o, y_o), is taken to be (0.5,0.5). The expected values and their variances for the direct calculation of the statistics are presented in Tables 2.1a and 2.1b.

Note that expected values are not known for harmonic means, but will be less than the expected values for the regular means.

The results of the simulations in a unit square with no clustering are presented in table 2.2. By comparing the observed statistics to the expected values, it can be seen that $\bar{r}^2$ and $\bar{d}^2$ are the only statistics for which the observed value closely approximates its expected value for all four values of

Table 2.1a
Expected Values of Statistics
in Unit Square

Statistic	Expected Mean	Variance (n=4)	Variance (n=6)	Variance (n=10)	Variance (n=20)
d	0.3802	0.00553	0.00369	0.00221	0.00111
d^2	0.1667	0.00278	0.00185	0.00111	0.00056
i	0.5774	0.01528	0.00833	0.00426	0.00189
$\bar{i}^2$	0.3333	0.02037	0.01111	0.00568	0.00251

Table 2.1b
Expected Values of Nearest Neighbor Statistic
in Unit Square

n	Expected Mean	Variance
4	0.2500	0.00425
6	0.2041	0.00189
10	0.1581	0.00068
20	0.1118	0.00017

n. Observed values for $\bar{d}$ slightly overestimate the expected, and $\bar{i}$ is always smaller than its expected value. The level of overestimation for $\bar{d}$ and d^2 does not appear to be dependent on n, as no trend is apparent.

The observed value for mean nearest-neighbor distance, r, is greater than the expected for all levels of n, although the amount of overestimation is dependent on n, as can be seen in figure 2.1.

Comparison of Jackknife and Bootstrap estimates to observed values shows that Jackknife is identical to the observed for all statistics except the harmonic means, for which the Jackknife estimate is always greater than the observed value. This relationship also appears related to the number of points, as can be seen in figures 2.2 and 2.3. The Jackknife variance is also identical to the observed variance, except for harmonic means, where the Jackknife variances are always less than the observed. Bootstrap estimates of the $\bar{d}$ statistics (d, d^2, and $\bar{d}_{harmonic}$) closely approximate the observed

values for mean and variance. Bootstrap estimates of $\bar{I}$ and $\bar{I}^2$, however, underestimate the observed values, and overestimate the observed values for $\bar{I}_{harmonic}$. This appears dependent on the number of points also, as can be seen in figures 2.4 and 2.5.

Table 2.2
Results of Monte Carlo Simulation Series 1
Unit Square, No Clustering

Stat.	Meth.	4 Points in Area		6 Points in Area		10 Points in Area		20 Points in Area	
		Mean	Var	Mean	Var.	Mean	Var.	Mean	Var.
$\bar{d}$	obs	0.3833	0.0052	0.3846	0.0032	0.3823	0.0020	0.3834	0.0010
	boot	0.3832	0.0052	0.3846	0.0032	0.3824	0.0020	0.3833	0.0010
	jack	0.3833	0.0052	0.3846	0.0032	0.3823	0.0020	0.3834	0.0010
$\bar{d}_{harm}$	obs	0.3307	0.0099	0.3226	0.0075	0.3096	0.0056	0.2999	0.0034
	boot	0.3306	0.0099	0.3226	0.0075	0.3097	0.0056	0.2998	0.0034
	jack	0.3392	0.0086	0.3269	0.0069	0.3110	0.0053	0.3004	0.0034
$\bar{d}^2$	obs	0.1674	0.0029	0.1680	0.0018	0.1665	0.0011	0.1674	0.0005
	boot	0.1674	0.0029	0.1680	0.0018	0.1665	0.0011	0.1673	0.0005
	jack	0.1674	0.0029	0.1680	0.0018	0.1665	0.0011	0.1674	0.0005
$\bar{l}$	obs	0.5196	0.0150	0.5236	0.0072	0.5195	0.0039	0.5215	0.0016
	boot	0.3898	0.0084	0.4364	0.0050	0.4676	0.0032	0.4954	0.0015
	jack	0.5196	0.0150	0.5236	0.0072	0.5195	0.0039	0.5215	0.0016
$\bar{l}_{harm}$	obs	0.3932	0.0181	0.3713	0.0092	0.3520	0.0046	0.3424	0.0015
	boot	0.5246	0.0324	0.4459	0.0133	0.3914	0.0057	0.3606	0.0017
	jack	0.4302	0.0151	0.3820	0.0081	0.3543	0.0042	0.3428	0.0015
$\bar{l}^2$	obs	0.3324	0.0204	0.3360	0.0103	0.3313	0.0058	0.3336	0.0024
	boot	0.2494	0.0115	0.2800	0.0072	0.2983	0.0047	0.3169	0.0022
	jack	0.3324	0.0204	0.3360	0.0103	0.3313	0.0058	0.3336	0.0024
r	obs	0.3214	0.0098	0.2512	0.0038	0.1839	0.0012	0.1242	0.0003
	boot	0.3215	0.0098	0.2513	0.0038	0.1839	0.0012	0.1242	0.0003
	jack	0.3214	0.0098	0.2512	0.0038	0.1839	0.0012	0.1242	0.0003

where:

$\bar{d}$ is the average distance to a fixed point
$\bar{d}_{harm}$ is the harmonic average distance to a fixed point
$\bar{d}^2$ is the average squared distance to a fixed point

$\bar{l}$ is the average interpoint distance
$\bar{l}_{harm}$ is the harmonic average interpoint distance
$\bar{l}^2$ is the harmonic average squared interpoint distance

r is the average nearest neighbor distance

obs refers to the observed value by direct calculation
boot refers to values generated by the Bootstrap method
jack refers to values generated by the Jackknife method

In the second series of simulations, an artificial cluster was established in the center of the unit square. The cluster was established in a square that included 20% of the total area, and was established in such a way that 3 times as many cases occur inside the square as outside. For the first two simulation series (with n=4, and n=6) this means only one point was included in the "cluster".

Expected values were not calculated for these simulations. Such values are complex to compute, and are beyond the scope of this dissertation. The degree to which observed values in this simulation deviate from the observed and expected values for the first series will reflect how well these statistics detect clustering; statistically different values from the first set of simulations should reflect clustering.

The results of these simulations of clustering in a unit square are presented in table 2.3. Jackknife and Bootstrap estimates relate to the observed values in the same way as in the first series of simulations, with the Jackknife exactly estimating the observed value, except for harmonic means, and Bootstrap closely estimating the observed distance to fixed point statistics (d, d^2 and $d_{harmonic}$), and nearest neighbor distance, r, underestimating $\bar{i}$ and $\bar{i}^2$, while overestimating $\bar{i}_{harmonic}$.

Nearest neighbor distances again decrease with increasing n, as do $\bar{i}$ and $\bar{i}^2$. The distance to fixed point statistics appear to have random variation over n.

In general, the observed distance to fixed point statistics were all less than their counterparts in the simulations with no clustering, as were the average interpoint distance statistics (i, i^2, and $i_{harmonic}$). The nearest neighbor statistics were all less than their counterparts in the first series. Figure 2.6 shows the ratio of observed statistics with clustering to that without

clustering for different values of n.

The variances of Jackknife, Bootstrap and directly calculated estimates are comparable for the nearest neighbor, $\bar{d}$ and $\bar{d}^2$ statistics. For the harmonic statistics, the Jackknife method yields larger variances than either the direct or the Bootstrap methods. With the interpoint statistics ($\bar{I}$, $\bar{I}^2$ and $\bar{I}_{harmonic}$), the Bootstrap method yields smaller variances, consistent with the underestimation this method introduces, as discussed above. For all statistics, variance decreases when the number of points in the analysis increases.

Table 2.3
Results of Monte Carlo Simulation Series 2
Unit Square, 3x Clustering in 20% Central Square

Stat.	Meth.	4 Points in Area		6 Points in Area		10 Points in Area		20 Points in Area	
		Mean	Var	Mean	Var.	Mean	Var.	Mean	Var.
$\bar{d}$	obs	0.3439	0.0014	0.3435	0.0009	0.3086	0.0005	0.3096	0.0003
	boot	0.3440	0.0014	0.3434	0.0009	0.3086	0.0005	0.3096	0.0003
	jack	0.3439	0.0014	0.3435	0.0009	0.3086	0.0005	0.3096	0.0003
$\bar{d}_{harm}$	obs	0.2359	0.0045	0.2286	0.0035	0.1961	0.0022	0.1911	0.0015
	boot	0.2360	0.0045	0.2286	0.0036	0.1959	0.0023	0.1911	0.0012
	jack	0.2492	0.0035	0.2343	0.0031	0.1979	0.0021	0.1917	0.0016
$\bar{d}^2$	obs	0.1531	0.0009	0.1529	0.0006	0.1290	0.0003	0.1299	0.0002
	boot	0.1532	0.0010	0.1529	0.0006	0.1290	0.0003	0.1299	0.0002
	jack	0.1531	0.0009	0.1529	0.0006	0.1290	0.0003	0.1299	0.0002
$\bar{l}$	obs	0.5032	0.0079	0.4983	0.0040	0.4491	0.0014	0.4479	0.0006
	boot	0.3776	0.0045	0.4152	0.0028	0.4041	0.0012	0.4255	0.0006
	jack	0.5032	0.0079	0.4983	0.0040	0.4491	0.0014	0.4479	0.0006
$\bar{l}_{harm}$	obs	0.3778	0.0129	0.3414	0.0072	0.2770	0.0026	0.2623	0.0010
	boot	0.5036	0.0230	0.4099	0.0028	0.3079	0.0033	0.2762	0.0011
	jack	0.4175	0.0105	0.3535	0.0064	0.2803	0.0025	0.2628	0.0010
$\bar{l}^2$	obs	0.3059	0.0102	0.3054	0.0054	0.2577	0.0018	0.2593	0.0008
	boot	0.2295	0.0059	0.2545	0.0038	0.2319	0.0015	0.2454	0.0007
	jack	0.3059	0.0102	0.3054	0.0054	0.2577	0.0018	0.2593	0.0008
r	obs	0.3029	0.0080	0.2352	0.0040	0.1580	0.0012	0.1001	0.0002
	boot	0.3030	0.0080	0.2351	0.0040	0.1580	0.0012	0.1001	0.0002
	jack	0.3030	0.0080	0.2352	0.0040	0.1580	0.0012	0.1001	0.0002

where:

$\bar{d}$ is the average distance to a fixed point
$\bar{d}_{harm}$ is the harmonic average distance to a fixed point
$\bar{d}^2$ is the average squared distance to a fixed point

$\bar{l}$ is the average interpoint distance
$\bar{l}_{harm}$ is the harmonic average interpoint distance
$\bar{l}^2$ is the harmonic average squared interpoint distance

r is the average nearest neighbor distance

obs refers to the observed value by direct calculation
boot refers to values generated by the Bootstrap method
jack refers to values generated by the Jackknife method

2.4. Zip Code Region Analysis

While the analysis of unit square data is useful for understanding the basic properties of these spatial statistics, such a construct is, almost by definition, artificial. The simulations were next repeated within a postal zip code region. Two series of simulations were performed, again with and without clustering of the data. The first simulation series reflected the behavior of the spatial statistics when cases are uniformly distributed in space, and the second series reflected the properties of the statistics given a non-uniform distribution in space. This second series may show how the statistics perform when cases cluster in a neighborhood.

The geographic region used for these simulations is a zip code region within Santa Clara County, California. The region is approximately 20 square miles in area, and its boundaries are defined by a highway and major roads. During 1983-1985, 1767 live births occurred in this zip code region, of which 40 had one or more birth defects, yielding a three-year crude birth defects prevalence proportion of 22.6/1,000. While the assumption of uniform distribution of live births in the region is not strictly correct, it is necessary to retain this assumption in order to evaluate the performance of the statistics under the null hypothesis.

Points were randomly placed in the region as for the first simulation series. To place a point in the region, a square enclosing the region was first defined. A random point was then selected and it was determined whether it lay inside or outside the boundary of the zip code area. If it lay outside, it was discarded and another point was selected. A FORTRAN program was used to accomplish point selection [16].

The expected values of r, the nearest neighbor distance, for the zip code region are shown in table 2.4. Results from four simulations (n=4,6,10,

and 20) in the zip code region with no clustering are shown as table 2.5. No expected values were calculated for the fixed point distance measures, or for the interpoint distance measures.

Table 2.4
Expected Values for
Nearest Neighbor Distance, r
Study Region Analysis

Number of Points	Expected r	Variance
4	1.11	0.0833
6	0.90	0.0370
10	0.70	0.0133
20	0.50	0.0033

The data in table 2.5 show that the observed value of r exceeds the expected value for all four values of n. As in the unit square, the nearest neighbor and interpoint distance statistics decrease with increasing n. The distance to a fixed point statistics do not vary significantly over n.

The Jackknife method performed the same way in the study area as in the unit square, approximating the observed value for all statistics except the harmonic means. The Bootstrap method also performed as in the unit square analyses, approximating the distance to a fixed point statistics and nearest neighbor, but significantly overestimating $\bar{i}$ and $\bar{i}^2$, and underestimating $\bar{i}_{harmonic}$. The variance of the Bootstrap statistics are somewhat smaller than the variance of the observed statistics for all except harmonic statistics and nearest neighbor.

Table 2.5
Results of Monte Carlo Simulation Series 3
Zip Code Region, No Clustering

Stat.	Meth.	4 Points in Area		6 Points in Area		10 Points in Area		20 Points in Area	
		Mean	Var	Mean	Var.	Mean	Var.	Mean	Var.
$\bar{d}$	obs	2.0047	0.2022	2.0358	0.1320	2.0124	0.0796	2.0186	0.0407
	boot	2.0039	0.2021	2.0366	0.1325	2.0127	0.0798	2.0183	0.0408
	jack	2.0047	0.2022	2.0358	0.1320	2.0124	0.0796	2.0186	0.0407
$\bar{d}_{harm}$	obs	1.6105	0.3187	1.5995	0.2566	1.5099	0.1678	1.4469	0.1002
	boot	1.6108	0.3187	1.6007	0.2570	1.5102	0.1682	1.4469	0.1002
	jack	1.6719	0.2845	1.6275	0.2386	1.5214	0.1611	1.4521	0.0963
$\bar{d}^2$	obs	4.8646	3.4334	4.9533	2.2547	4.8664	1.3939	4.9020	0.7122
	boot	4.8611	3.4294	4.9564	2.2678	4.8679	1.4006	4.9012	0.7128
	jack	4.8646	3.4334	4.9533	2.2547	4.8664	1.3939	4.9020	0.7122
$\bar{l}$	obs	2.3246	0.2885	2.3295	0.1536	2.3230	0.0802	2.3283	0.0334
	boot	1.7429	0.1627	1.9404	0.1075	1.7680	0.1061	1.6053	0.0341
	jack	2.3246	0.2885	2.3295	0.1536	2.3230	0.0802	2.3283	0.0334
$\bar{l}_{harm}$	obs	1.7819	0.3549	1.6555	0.1925	1.5893	0.0855	1.5255	0.0306
	boot	2.3790	0.6323	1.9885	0.2787	1.7680	0.1061	1.6053	0.0341
	jack	1.9363	0.2939	1.7049	0.1701	1.6008	0.0812	1.5275	0.0301
$\bar{l}^2$	obs	6.6123	7.6975	6.6401	4.3899	6.6241	2.3444	6.6536	1.0087
	boot	4.9565	4.3345	5.5303	3.0673	5.9636	1.8931	6.3194	0.9078
	jack	6.6123	7.6975	6.6401	4.3899	6.6241	2.3444	6.6536	1.0087
r	obs	1.4418	0.1976	1.1165	0.0792	0.8280	0.0232	0.5533	0.0050
	boot	1.4424	0.1978	1.1164	0.0795	0.8281	0.0232	0.5534	0.0050
	jack	1.4418	0.1976	1.1165	0.0792	0.8280	0.0232	0.5533	0.0050

where:

$\bar{d}$ is the average distance to a fixed point
$\bar{d}_{harm}$ is the harmonic average distance to a fixed point
$\bar{d}^2$ is the average squared distance to a fixed point

$\bar{l}$ is the average interpoint distance
$\bar{l}_{harm}$ is the harmonic average interpoint distance
$\bar{l}^2$ is the harmonic average squared interpoint distance

r is the average nearest neighbor distance

obs refers to the observed value by direct calculation
boot refers to values generated by the Bootstrap method
jack refers to values generated by the Jackknife method

A second series of simulations was performed in the study area, with the addition of the clustering of points within a defined area. For convenience, although additional points were added to the maps in the clustering simulations, the simulations will be referred to by the number of points initially in the map.

In the zip code region simulation, clustering was established by addition of more points to a circular subregion 2 miles in diameter, located in the western portion of the zip code area (figure 2.7). The location and size of this clustering were arbitrarily selected. The cluster was created to have a relative risk of 3.0, compared to the rest of the zip code region, assuming a uniform distribution in the region of the population at risk.

The results of these clustered simulations in the study area are presented in table 2.6. Consistent with the other simulations, the Jackknife estimates of all but the harmonic statistics are identical with the mean value calculated directly. Bootstrap estimates and their variances are also consistent with the directly estimated means of the distance to a fixed point and nearest neighbor statistics, and underestimate $\bar{r}$ and $\bar{r}^2$ due to sampling with replacement, which results in sampling the same point more than once in many Bootstrap samples.

Clustering becomes evident with all statistics when approximately 10 points have been included in the area. In this scenario, five of the points on the map will occur in the cluster region. In the scenarios with four or six points on the map, the number of clustered points is again too small to be detected by any of the statistics when compared to values obtained in the unclustered simulations. One cannot directly compare the simulations with and without clustering in the zip code region for the nearest neighbor and average interpoint distance statistics, because these statistics are dependent

on the number of points on the map. As noted, additional points were added to these maps to simulate clustering. Thus, while all statistics appear to detect clustering when one starts with 10 points, in fact the ratio of clustered:non-clustered values for the $\bar{r}$ and r statistics are smaller than they would have been, had the correct number of points been used in the comparison population.

For the nearest neighbor statistic, values in the clustered simulation are actually larger than the expected values tabulated in table 2.4 for n less than 10, and approximately equal to the expected values for the 10 and 20 point simulations.

Table 2.6
Results of Monte Carlo Simulation Series 4
Zip Code Region, 3x Clustering in 16% Area Circle

Stat.	Meth.	4 Points in Area		6 Points in Area		10 Points in Area		20 Points in Area	
		Mean	Var	Mean	Var.	Mean	Var.	Mean	Var.
$\bar{d}$	obs	2.0069	0.2215	2.0366	0.1296	1.7035	0.0477	1.7000	0.0266
	boot	2.0067	0.2220	2.0366	0.1298	1.7038	0.0788	1.7001	0.0267
	jack	2.0069	0.2215	2.0366	0.1296	1.7035	0.0477	1.7000	0.0266
$\bar{d}_{harm}$	obs	1.6324	0.3314	1.5922	0.2361	0.9778	0.0721	0.9351	0.0425
	boot	1.6324	0.3321	1.5922	0.2357	0.9777	0.0723	0.9346	0.0436
	jack	1.6889	0.3033	1.6210	0.2191	0.9862	0.0700	0.9375	0.0420
$\bar{d}^2$	obs	4.8696	3.8856	4.9696	2.2984	3.9075	0.8047	3.8987	0.4394
	boot	4.8699	3.8946	4.9695	2.3000	3.9099	0.8103	3.8986	0.4413
	jack	4.8696	3.8856	4.9696	2.2984	3.9075	0.8047	3.8987	0.4394
$\bar{l}$	obs	2.3008	0.2920	2.3464	0.1624	2.1874	0.0547	2.1773	0.0267
	boot	1.7255	0.1648	1.9548	0.1124	2.0185	0.0464	2.0929	0.0246
	jack	2.3008	0.2920	2.3464	0.1624	2.1874	0.0547	2.1773	0.0267
$\bar{l}_{harm}$	obs	1.7485	0.3518	1.6759	0.1962	1.4245	0.0531	1.3910	0.0245
	boot	2.3316	0.6298	2.0113	0.2825	1.5439	0.0627	1.4472	0.0266
	jack	1.9103	0.2905	1.7232	0.1743	1.4307	0.0518	1.3922	0.0243
$\bar{l}^2$	obs	6.5050	7.9044	6.7374	4.6524	5.9466	1.4554	5.9052	0.7037
	boot	4.8785	4.4541	5.6122	3.2223	5.4881	1.2391	5.6755	0.6501
	jack	6.5050	7.9044	6.7374	4.6524	5.9466	1.4554	5.9052	0.7037
r	obs	1.4201	0.1878	1.1257	0.0762	0.6704	0.0138	0.4685	0.0033
	boot	1.4199	0.1887	1.1256	0.0761	0.6703	0.0139	0.4683	0.0033
	jack	1.4201	0.1878	1.1257	0.0762	0.6704	0.0138	0.4685	0.0033

where:

$\bar{d}$ is the average distance to a fixed point
$\bar{d}_{harm}$ is the harmonic average distance to a fixed point
$\bar{d}^2$ is the average squared distance to a fixed point

$\bar{l}$ is the average interpoint distance
$\bar{l}_{harm}$ is the harmonic average interpoint distance
$\bar{l}^2$ is the harmonic average squared interpoint distance

r is the average nearest neighbor distance

obs refers to the observed value by direct calculation
boot refers to values generated by the Bootstrap method
jack refers to values generated by the Jackknife method

2.5. Discussion

2.5.1. Unit Square

In the unit square analysis, several properties of the spatial statistics used become clear. First, both the nearest neighbor and average interpoint distance statistics decrease with increasing n. This is as expected, as the more points in a confined space, the smaller the distance must be between those points. While this property is accounted for in the expected values of r, the expectations for $\bar{r}$ and $\bar{r}^2$ do not take into account the number of points in the sample, and thus the expectation for $\bar{r}$, as calculated, does not vary over n.

Variation in the nearest neighbor distance from its expected value is probably due primarily to boundary effects. As described in chapter 1, one issue with the use of the nearest neighbor statistic is that, when n is small, the nearest point in a uniform distribution may be outside the boundary of the area. This property exists at any n for points near the edge of the area under consideration, but is particularly critical when n is small, as the distances are relatively large. While this effect is important when comparing the observed r to the theoretically calculated expected value, it is possible an overriding concern would be the assumption of a uniform distribution of the underlying population at risk. The use of a randomization or permutation technique, as will be demonstrated in chapter 4, can partially control for both of these potential biases.

The values of $\bar{d}$ and $\bar{d}^2$ are consistent with the expected values, and although a slight over-estimation appears to exist for $\bar{d}$, it is not statistically significant, and is probably due to random variation in the location of the sample points. The squared average interpoint distance ($\bar{r}^2$) is consistent

with its expected value in all cases. The average interpoint distance ($\bar{i}$), however, appears to be underestimated. As $\bar{i}^2$ is so close to its expected value, it is probable that the difference in observed and expected values for $\bar{i}$ is due to the approximation of the expected value as the square root of the expected value of $\bar{i}^2$. Again, the underestimation is not statistically significant.

The use of the Jackknife for the non-harmonic statistics in a unit square with no clustering is clearly not necessary, as the Jackknife adds no information to what can be learned from the observed value. The Jackknife exactly estimates the mean, which is to be expected since, on reflection, the Jackknife mean is mathematically equivalent to the observed mean. With respect to the harmonic means, the Jackknife estimate has a smaller variance than the observed harmonic mean, and thus is probably a slightly better estimator, although the difference in variances is very small.

The Bootstrap method performs extremely well for the distance to a fixed point statistics, and for the nearest neighbor distance, but is a poor estimator of the average interpoint distances. This is due to the fact that in the Bootstrap sampling scheme, the same point may be selected twice or more frequently in the same sample, or not at all. If n=4, each Bootstrap sample may contain one of the four points once, twice, three or four times, or not at all. If a point appears two or more times, its interpoint distance to itself will be 0, thus causing the mean to underestimate the true value. In harmonic means, the interpoint distance of 0 will be treated as infinite, as $\frac{1}{0}$ is undefined. When this situation arose, the estimates of $\bar{i}$, $\bar{i}^2$, and $i_{harmonic}$ for that replication were not used in the calculation of the overall harmonic means. As n increases, the probability of a single point appearing more than once in a Bootstrap sample decreases, and as such the Bootstrap

estimate performs better with increasing n. As such, the Bootstrap method cannot be effectively used for interpoint distance statistics. It should be noted that for the other statistics, the Bootstrap estimates had, by far, the lowest variances. This is a reflection on the number of Bootstrap samples taken, and thus is not directly comparable to the variances of the directly estimated values.

In the second series of simulations, with clustering in the unit square, it becomes clear that all of the spatial statistics respond to the increased clustering, as all statistics are reduced in the clustered scenario, compared to the random scenario. In the nearest neighbor statistics, a reduction in the "boundary effect" is observed more rapidly with increasing n, probably as a result of the clustering in the center of the square, which effectively "pulls" the points away from the boundary, thereby reducing the boundary problem described above.

Comparing the $\bar{d}$ and $\bar{r}$ statistics observed in the clustered simulation to the expected values under no clustering in table 2.1, the only cluster significant at α=0.05 is the 20-point simulated cluster. None of the clusters were statistically significant for the $\bar{d}^2$, $\bar{r}^2$, or τ statistics.

The clustering of the 6 point maps was not distinguishable from the 4 point map with any statistics. The six point map was not significantly different from the four-point map, since both maps implied only one or two additional points in the cluster. Such small changes are not likely to be differentiated by these statistics.

2.5.2. Study Region

As noted, the observed values of the nearest neighbor distance in the zip code region analysis were also greater than the expected values for the simulations with no clustering. This is probably the result of a continued boundary-effects problem, as described above. The harmonic mean interpoint distance, $\bar{r}_{harmonic}$, was much smaller than in the unit square simulations. All other statistics are larger due to the difference in scale. As the map is not one unit wide, these distances would be expected to be larger.

The nearest neighbor distance and the average distance to a fixed point are the most sensitive statistics for detecting clusters, as can be seen in figure 2.8. This is consistent with the unit square, although the comparisons made in this figure are not strictly correct, since the number of points in the clustered maps are different than in the non-clustered map, which will bias the ratio of the nearest neighbor distance and average interpoint distance statistics to be smaller than they should be. Since the size of the zip code region is large, the effect of sample size on these statistics is small, as can be seen in table 2.5. The nearest neighbor distance in the clustered investigation does not successfully measure clustering when compared to the expected value, however.

The size of the cluster will also affect the ability of the statistics to detect clustering. The demonstration cluster, as seen in figure 2.7, is large, and thus the points added to simulate clustering will still be dispursed. It should be emphasized that the "clustering" with four or six points in the initial map would not be measurable, since the number of points added to generate an elevated relative risk is small.

2.6. Summary

Each of the spatial statistics evaluated in this chapter has its own strengths and weaknesses for use in the investigation of clusters. The nearest neighbor statistic, while probably the most commonly used of the spatial statistics described here, is actually not very sensitive to the clustering introduced in these simulations, as demonstrated in both the unit square, and in the study region. Indeed, of the seven statistics evaluated, the nearest neighbor was the least sensitive, when compared to its expected value, due to the strong effect the boundary of the study area has on its value. The average interpoint distance and the distance to a fixed point each measure slightly different aspects of clustering, and are appropriate for different types of clusters; the $\bar{d}$ statistics being more appropriate when a point source is hypothesized as the cause of clustering, while the $\bar{I}$ statistics are perhaps more appropriate for clusters where no causal hypothesis has been made. Both statistics are moderately sensitive at detecting clusters when compared to their expected value, but the cluster must include a sufficient number of cases in order for the difference to become statistically detectable.

No obvious benefit arises from the use of the Jackknife or Bootstrap techniques in either the unit square or zip code region. Indeed, the use of the Bootstrap method produces erroneous results when applied to average interpoint distances, as the sampling with replacement often results in selecting the same point more than once, making the Bootstrapped value of the $\bar{I}$ or $\bar{I}^2$ statistic smaller than it is actually expected to be, and conversely making the $\bar{I}_{harmonic}$ statistic correspondingly larger than it is expected to be. Since the Jackknife method perfectly estimates mean values, its use for these statistics, all of which are means, is redundant and unwarranted.

The behavior of these statistics in a unit square was predictive of their behavior in the zip code region. The disadvantage of the zip code region is that expected values for the statistics could not be readily calculated. However, the unit square analysis showed that the simulated statistics were accurate estimates of the theoretically calculated expected values, and thus expected values in the zip code region can be approximated by the estimated values from the simulation in the zip code region without clustering.

Based on these results, the reanalysis of a spatial cluster, to be conducted in chapter 4, will make use of direct estimation, and will not use the Jackknife or Bootstrap methods for estimation. As there are approximately 15 children born with birth defects in the study region to be used in that chapter, the sample size should be sufficient for the detection of clustering, using these methods.

2.7. References

1. Knox, G. Detection of low intensity epidemicity: application to cleft lip and palate. *Br J Prev Soc Med.* 17: 121-127; 1963.
2. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220; 1967.
3. Miller, R. G. The jackknife—a review. *Biometrika.* 61: 1-15; 1974.
4. Tukey, J. W. Bias and confidence in not-quite large samples (abstract). *Ann Math Statist.* 29: 614; 1958.
5. Quenouille, M. H. Notes on bias in estimation. *Biometrika.* 43: 353-360; 1956.
6. Diaconis, P.; Efron, B. Computer-intensive methods in statistics. *Sci Am.* 248: 116-130; 1983.
7. Efron, B. Bootstrap Methods: another look at the Jacknife, the 1977 Rietz lecture. *Annals of Statistics.* 7: 1-26; 1979.
8. Efron, B. Nonparametric estimates of standard error: the jacknife the bootstrap and other methods. *Biometrika.* 68: 589-599; 1981.
9. Efron, B.; Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Statist.* 37: 36-48; 1983.
10. Koepsell, T. The bootstrap: examples of its application. *116th Annual Meeting of the American Public Health Association.* Boston, MA; Nov 13-17, 1988.
11. Abbott, R. D.; Carroll, R. J. Interpreting multiple logistic regression coefficients in prospective observational studies. *Am J Epidemiol.* 119: 830-836; 1984.

12. Weiss, S. T.; Munoz, A.; Stein, A. The relationship of blood lead to blood pressure in a longitudinal study of working men. *Am J Epidemiol.* 123: 800-808; 1986.

13. Mapleson, W. W. The use of GLIM and the bootstrap in assessing a clinical trial of two drugs. *Stat Med.* 5: 363-374; 1986.

14. Williams, P. T.; Fortmann, S. P.; Terry, R. B. Associations of dietary fat, regional adiposity, and blood pressure in men. *JAMA.* 257: 3251-3256; 1987.

15. Selvin, S.; Merrill, D.; Schulman, J.; Sacks, S.; Bedell, L.; Wong, L. Transformations of maps to investigate clusters of disease. *Soc Sci Med.* 26: 215-221; 1988.

16. Selvin, S. Unpublished computer program. 1989.

Figure 2.1
Ratio of Observed/Expected
Nearest Neighbor Distance, r

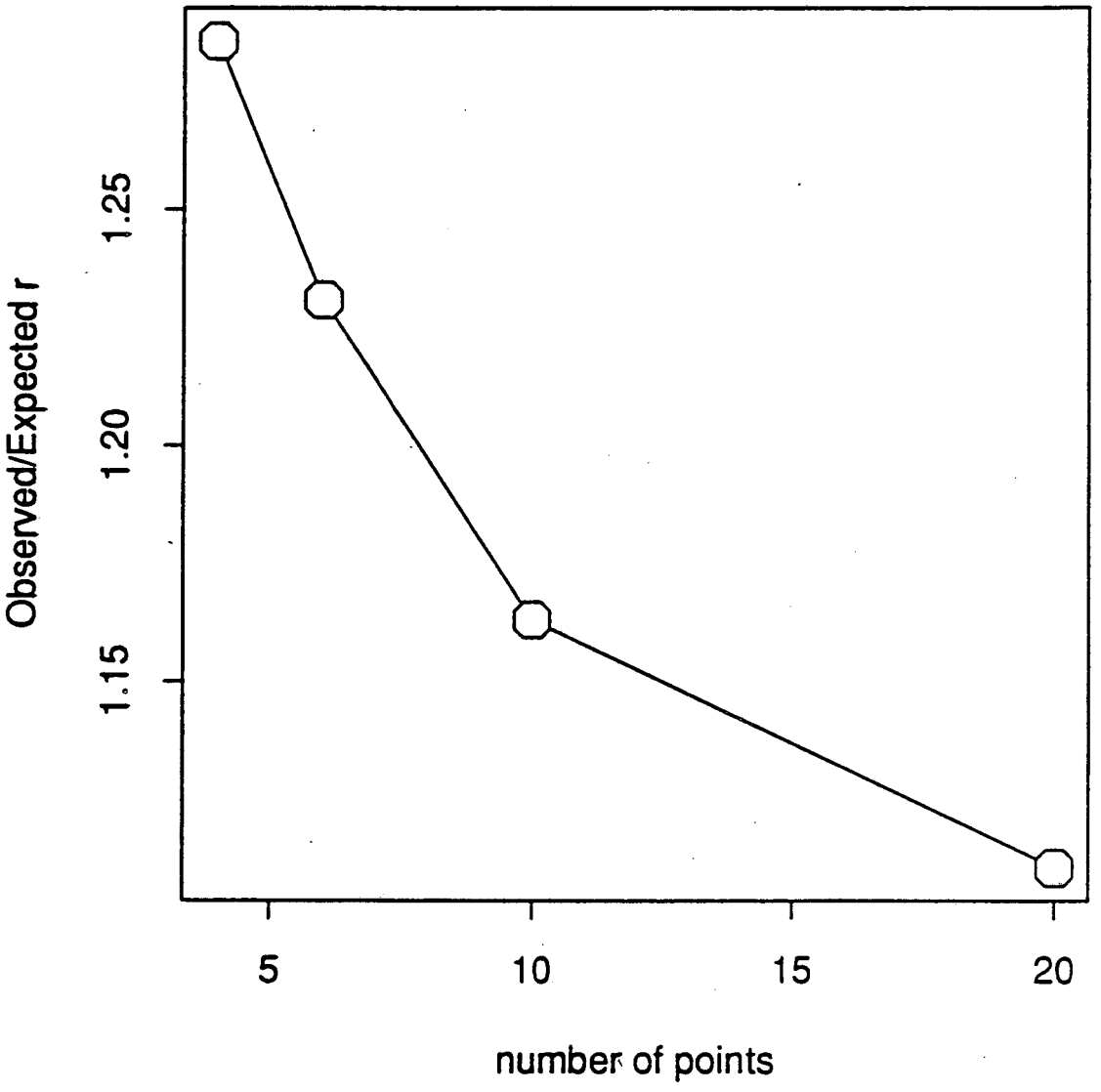


Figure 2.2
Ratio of Jacknife/Observed
Harmonic Average Distance to Fixed Point

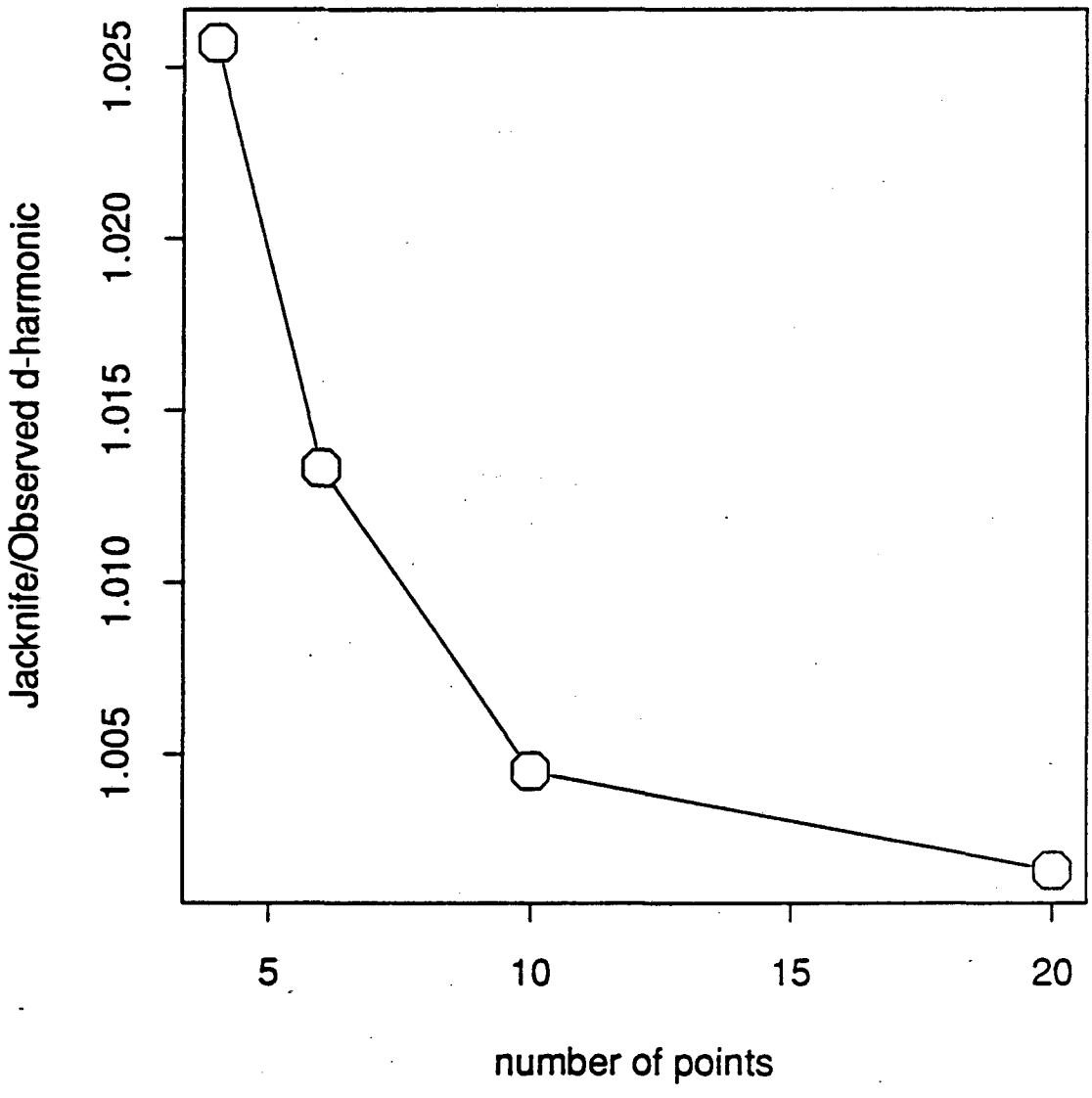


Figure 2.3
Ratio of Jacknife/Observed
Harmonic Average Interpoint Distance

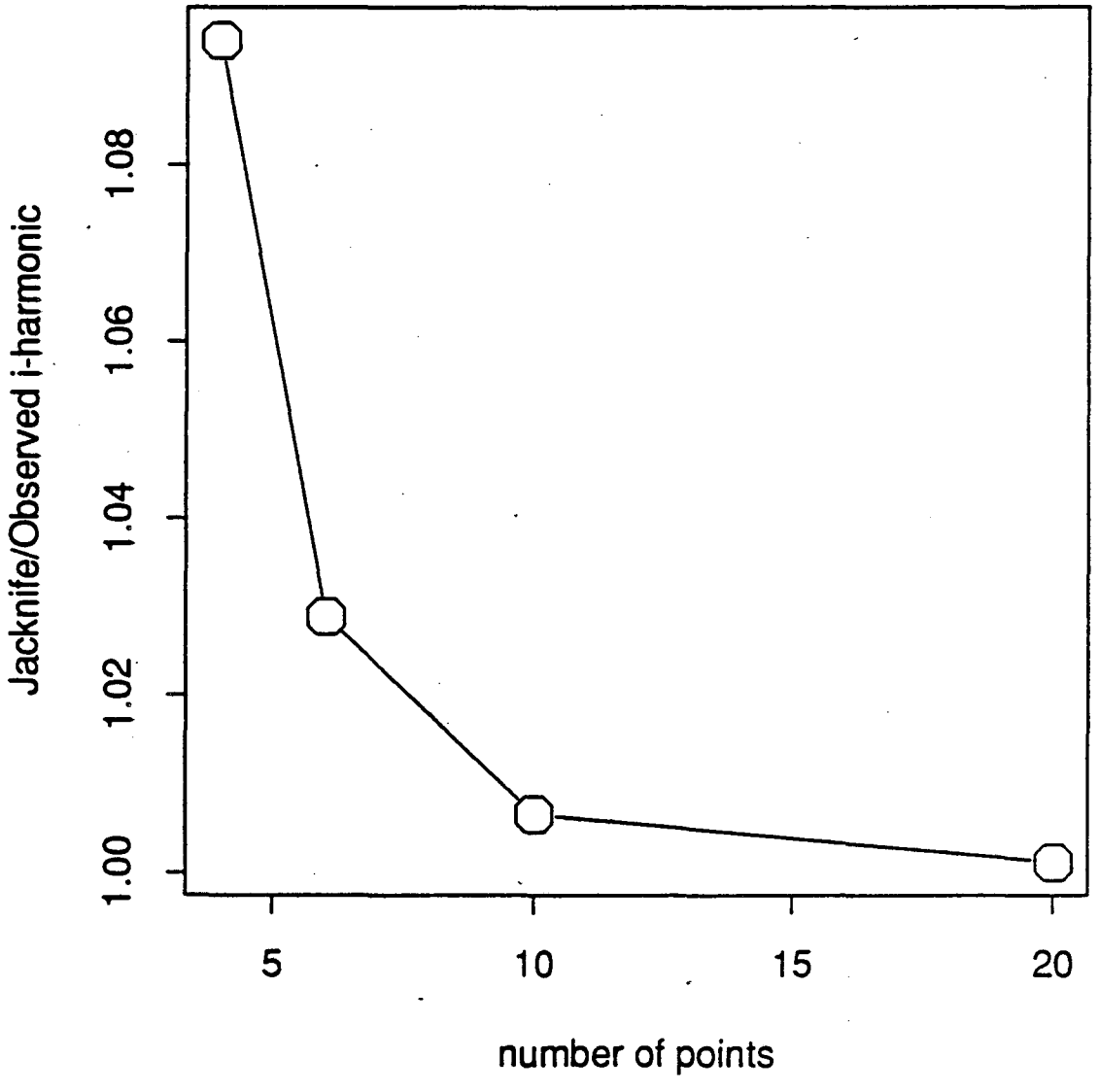


Figure 2.4
Ratio of Bootstrap/Observed
Average Interpoint Distance, i

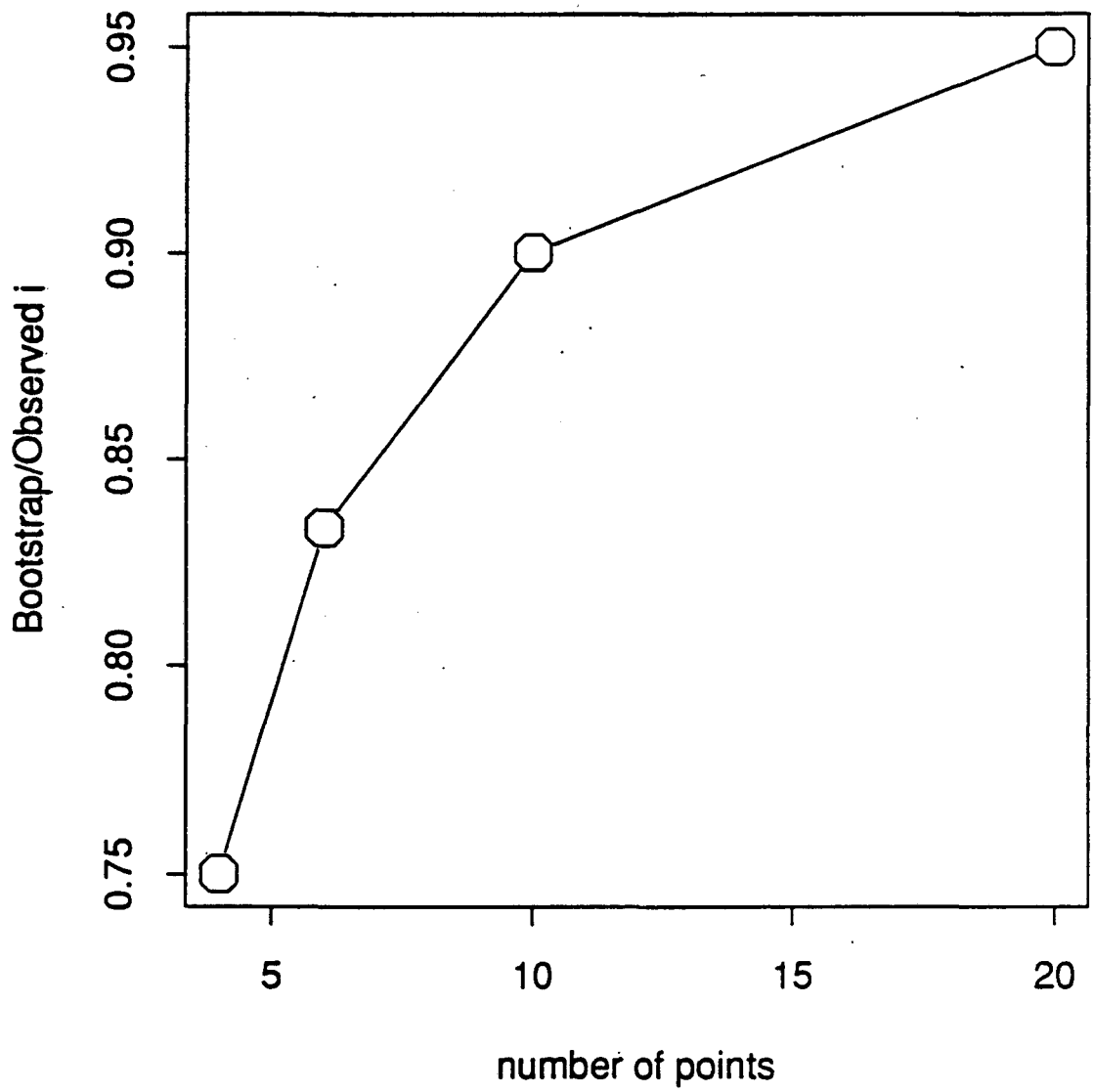


Figure 2.5
Ratio of Bootstrap/Observed
Harmonic Average Interpoint Distance

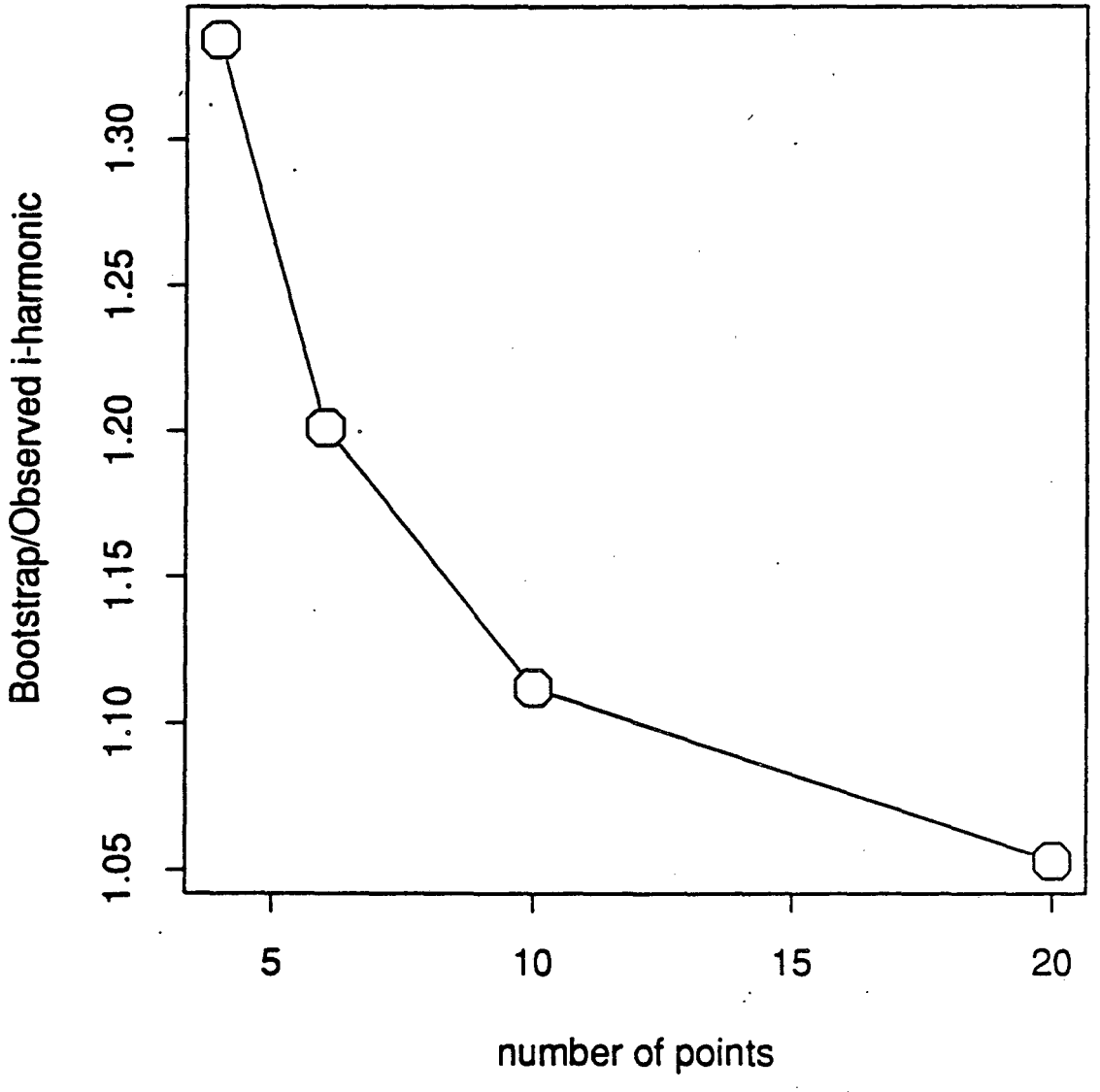


Figure 2.6
Ratio of Observed Statistics
Cluster:No Cluster, Unit Square

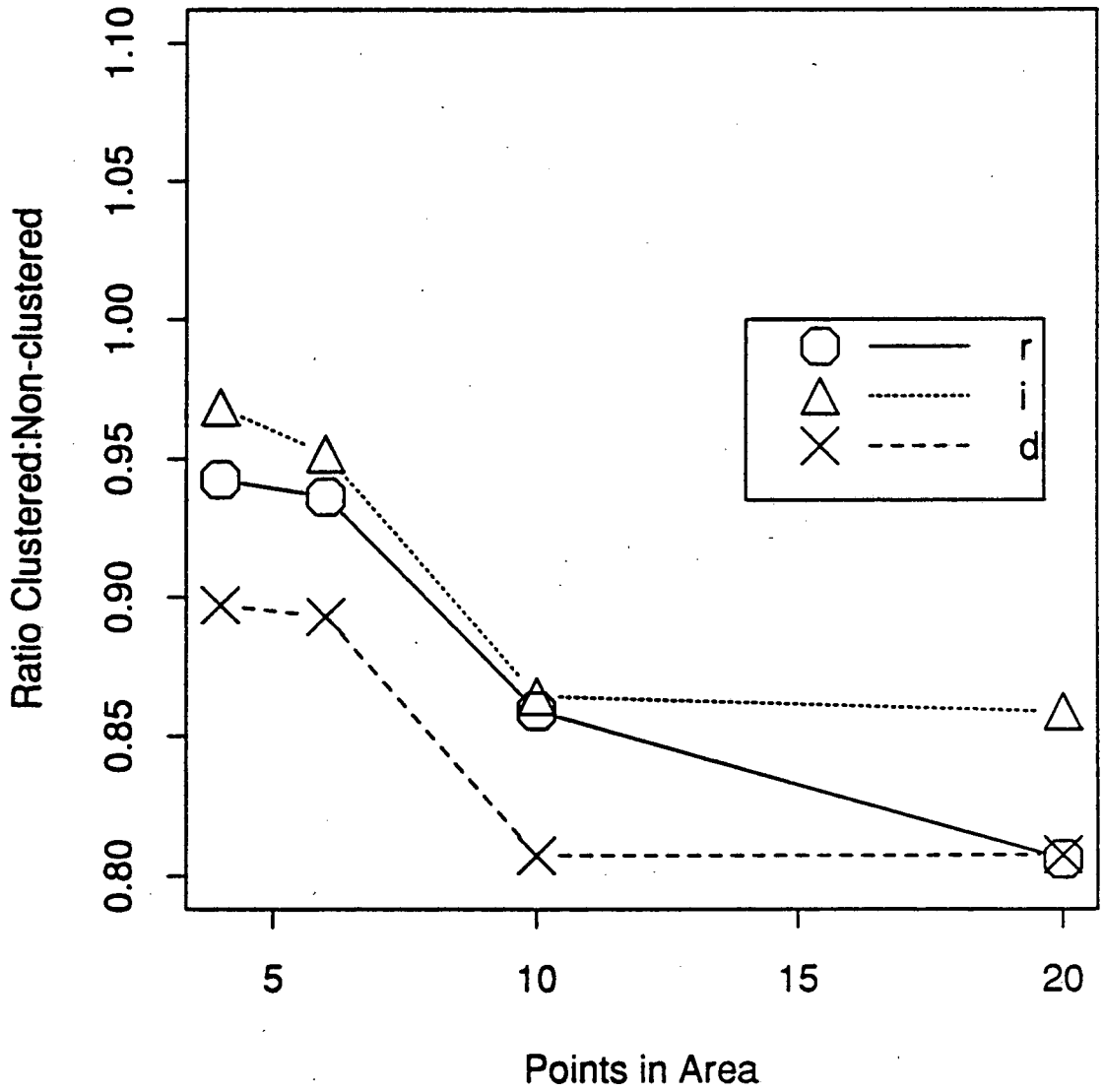


Figure 2.7
Zip Code Region with Cluster Area Shown
Santa Clara County

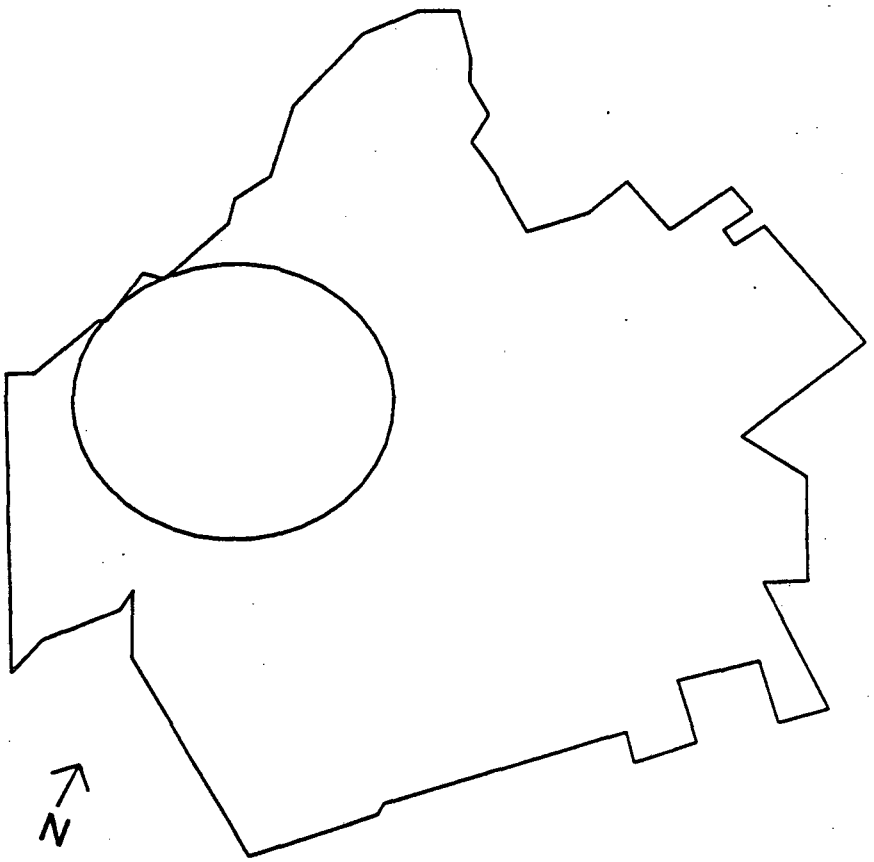
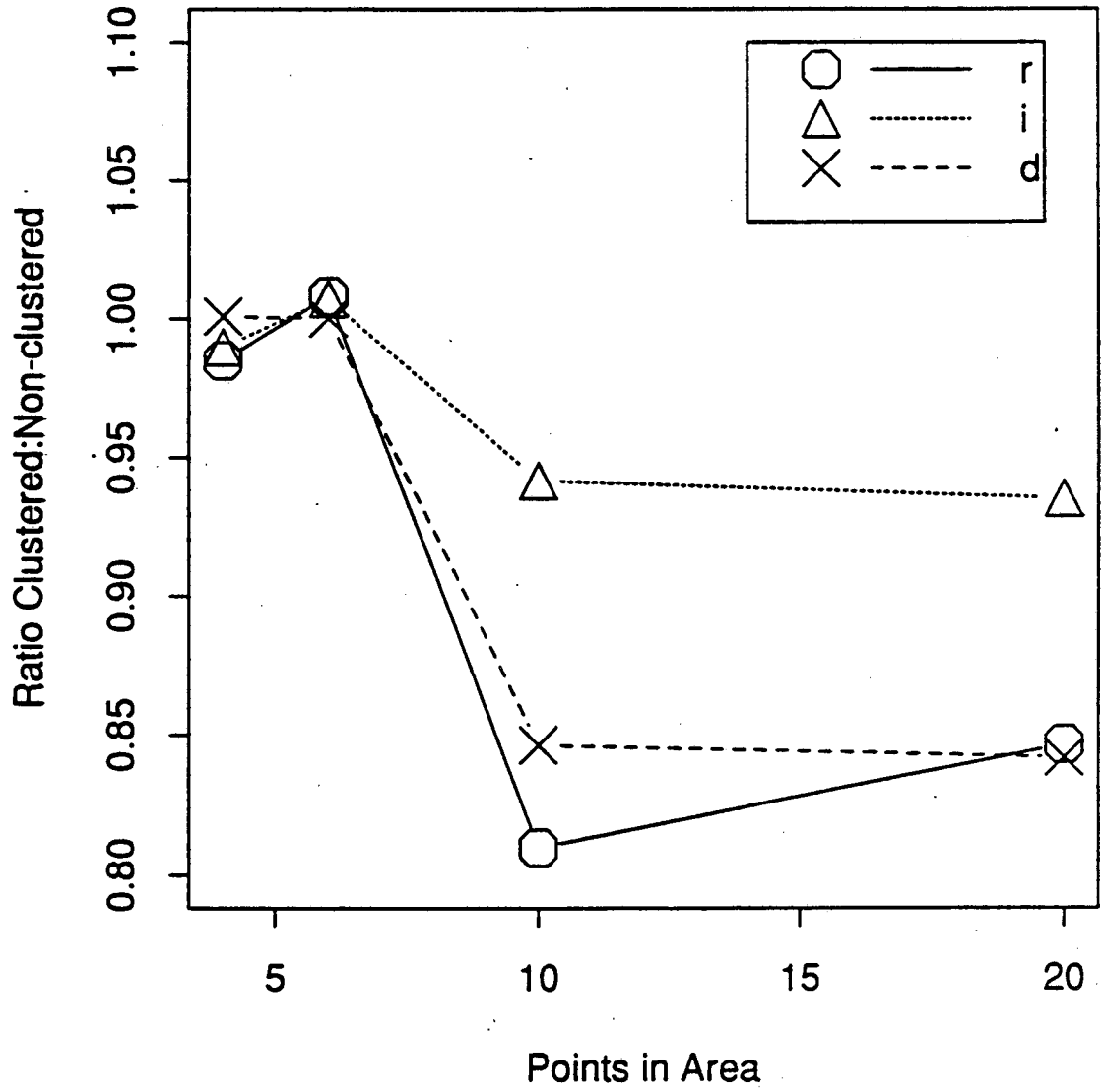


Figure 2.8
Ratio of Observed Statistics
Cluster:No Cluster, Zip Code Region



3. Map Comparison

3.1. Introduction

The investigation of spatial clustering in epidemiology has become an important tool for public health professionals responding to public concerns, often raised in response to actual or perceived environmental health risks. Recent reports of cancer and birth defects clusters (e.g. [1,2]) have underscored the need for adequate statistical methods for the investigation of such clusters. Many of the statistics used for the initial investigation of a cluster are based on an assumption that is often violated: that the underlying population at risk is uniformly distributed in the geographic region of concern. This assumption is made in lieu of determining the actual spatial distribution of this underlying population, which may be impossible to do, and in any case would involve considerable effort. Methods to control for the spatial distribution of the underlying population at risk have been introduced [3,4], but these methods involve the use of considerable computer resources that may not be available to public health departments, and in addition require that the analysis be performed using spatial boundaries, such as those of census tracts, that are often not available.

Often, the spatial distribution of a population at risk is not available (e.g., is not stored in electronic form). An example of such a population is that of live births. While electronic vital statistics files maintain some spatial data, such as zip code and possibly census tract, the precise address of residence at birth for live births is generally not maintained in electronic form. To make use of the spatial distribution of live births in a cluster investigation, one must either restrict analysis to comparison of rates by census tract or zip code region (a method that is generally not sensitive to

clustering, unless the cluster is truly dramatic), or one must manually abstract the addresses from the confidential portion of the birth certificate, which is not always available for review.

One alternative is to use some surrogate population, for which spatial location data are available, to represent the population at risk. Surrogate populations have been used extensively in epidemiology when information about the true population at risk is not available, or is difficult to obtain. (see for example [5-7]). To use such populations for cluster investigations, it is necessary to make the assumption that the surrogate population is distributed in space the same way as the true population at risk. One logical surrogate population to consider for cluster investigations of specific birth defects is the pooled population of all birth defects, regardless of type. This potential surrogate population is advantageous since data on children born with birth defects are available through registries [8,9], since such registries may maintain street address in electronic form, and since these registries are often the focal point for cluster investigations.

Confounding and selection bias also become an issue if a population with a particular genetic or socially-induced risk factor for producing children with defects lived in a small area. Such groupings can be easily constructed; a simple illustration might be that people with low socio-economic status (SES) may live in specific neighborhoods. The effect of such grouping of the referent population on the detection of a cluster depends on the location of the cluster with respect to the population that is grouped. If the cluster is located within such a population, it may not be detected as readily, due to the increased prevalence of defects within that area. In this situation, the clustering of the referent population may actually be of benefit, since it essentially controls for SES, or whatever confounding factor is causing the

grouping of the referent population, in the analysis. Spurious clustering, however, may also be observed if the neighborhood at high risk for the confounding factor (SES, in our example) is surrounded by neighborhoods at different risk due to the confounder. An example of this might be a low SES neighborhood bordering on middle class neighborhoods. Since the prevalence of disease in the low SES neighborhood is higher than that in the surrounding area, if a larger spatial unit, such as a census tract or zip code region, is used, the neighborhood may be falsely suspected of having a cluster of disease due to an environmental cause, rather than because of the confounding variable.

While the assumptions made for other epidemiologic investigations with respect to surrogate populations may be unsubstantiated, in many cases, it has not been determined whether their use in spatial analysis is viable. Such use would be beneficial as data on the spatial distribution of the underlying population at risk is often not available, while data on surrogate populations (such as in registries) may be readily available.

The focus of this chapter is to investigate whether the distribution of all birth defects can be applied as a surrogate for all live births in the investigation of clusters without incurring substantial bias. For the population of all birth defects to be used as a control population, it is necessary to show that this population is spatially not distinguishable from the population of live births; that is, that birth defects as a group are, at least approximately, randomly scattered in the population of live births with respect to space.

3.2. Methods

The study area for this investigation will be Santa Clara County, California. Santa Clara County encompasses 1,293 square miles [10], and incorporates a large urban center in San Jose, the county seat, and in the "Silicon Valley" region of the north and northeast. The county has a birth rate of 17.2/1000 per annum [11], of which approximately 26.25 per 1000 births are born with at least one birth defect (95% confidence interval [25.2,27.3]) [12]. During the years 1983-1985, 70,489 live births occurred in Santa Clara county [11, 13, 14], of which 1760 had one or more structural defects [12]. Data on birth defects in Santa Clara county are available from the California Birth Defects Monitoring Program [12] for the years 1983 to date, and data on address at birth were abstracted at the County vital statistics office.

Santa Clara county has a population of about 1,400,000, with an ethnic makeup of 71% white, 3% black, 8% asian, 17% hispanic and 1% other. Approximately 2% of Santa Clara County's population lives in rural areas. In 1986, the county had a 6% unemployment rate, and the major occupations were agriculture, manufacturing, service, and retail [15].

A random sample of 218 live births was randomly selected from a data set containing all live births with no birth defects, as defined by the California Birth Defects Monitoring Program (see below), in Santa Clara County for the years 1983, 1984 and 1985. The number sampled was selected to assure approximately 200 cases in the dataset after exclusions. As birth address is not recorded in electronic form on county vital statistics tapes, the addresses of these live births, as recorded on the birth certificate, were manually abstracted from the vital statistics files in Santa Clara county. Seventeen of the addresses were excluded from analysis. Seven records were not available from the county, and 10 addresses were either incorrect (address could

not be located in the county street index [16], on the county assessor's maps [17], or using a commercial street map book [18]), or were post office boxes that do not reflect the actual location of the mother during gestation. The remaining 201 live births were used in this analysis.

Birth defects were identified using the California Birth Defects Monitoring Program (CBDMP), an active surveillance system enacted by state law in California. The CBDMP identifies children born with structural birth defects, defined as those included in the British Pediatric Association (BPA) codes 740.0 - 759.9 that can be identified from 20 weeks gestation through the first year of life [8]. These data are collected from hospitals, genetic disease registries, physicians and other sources and are compiled for all counties in California, for live births, spontaneous abortions and still births after 20 weeks gestation, and for therapeutic and spontaneous abortions before 20 weeks gestation where the medical record indicates the presence of one or more structural defects. Surveillance by the CBDMP started in Santa Clara County in 1983.

A random sample of 240 children born with at least one defect was taken from this registry for the years 1983, 1984 and 1985. The number sampled was also selected to assure a minimum of 200 children with birth defects in the final dataset. Street address at birth was abstracted from the CBDMP database. Twenty-four of the records were excluded from analysis, due to incorrect address, address not representative of location of the mother during gestation, as above, or no birth address listed in the registry.

The addresses of the live births and of the children born with birth defects were then converted to spatial coordinates using a Summagraphics digitizing tablet connected to a microcomputer running commercially available software [19]. Addresses were located using the county assessor's

maps[17], in conjunction with a commercial street map book [18]. The county assessor's maps made use of the California Coordinate System of 1983, zone III, which uses a selected point (120° 30′ *W*, 36° 30′ *N*), as origin, and assigns coordinates based on feet in the X and Y direction from this arbitrary point [20]. Correction for the earth's curvature is not necessary so long as the points being digitized are in a limited geographic area. In all figures in this dissertation, the coordinate system has been removed to maintain confidentiality of the birth defect data.

A statistical comparison of the two samples was explored using several methods: the Kolmogorov-Smirnov test and the t-test, applied to a distance measure, Hotelling's T^2 test, applied to the coordinates of individual points, and nearest neighbor methods.

To accomplish the univariate statistical analysis, the distance of each point in both samples to a fixed point was calculated. For the purpose of this investigation, the fixed point used was determined by first pooling the samples of live births and birth defects, and then calculating the location of the centroid of the resulting population of 417 points, the "pooled centroid". A t-test, comparing the means of these two samples, was then performed. In addition, the Kolmogorov-Smirnov test, a non-parametric test that compares the cumulative frequency distributions was used [21].

Analysis of the spatial coordinates was performed using Hotelling's T^2 test [22]. Hotelling's T^2 is a multivariate t-test, and assesses whether the variation of the mean x and y coordinates is because of chance or to some systematic difference. Hotelling observed that T^2 was related to an F distribution with 2 and n-2 degrees of freedom according to equation 3.1 [22].

$$F_{2,n-2} = T^2 \times \frac{(n-2)}{2\,(n-1)} \qquad (3.1)$$

Since it was unclear whether the assumption of bivariate normality underlying the use of the F distribution was met by the present data, a permutation test was also performed. To accomplish this, points were randomly taken from the list of coordinates pooled from the samples of live births and birth defects and assigned to one of two groups, and the resulting T^2 value was calculated. This process was repeated 2000 times. The probability of the observed T^2 can then be compared to the 2000 values generated, and a "p-value" can be assigned.

The reasoning behind this process is intuitive: if live births and birth defects are distributed in the same way, then statistically it doesn't matter which group any given point is assigned to, since the only variation is random. We should, therefore, be able to assign the points to either group without affecting the T^2 value. In this particular case, there are $\frac{417!}{201!216!}$ possible combinations of these points, and to be truly rigorous, one would have to calculate the T^2 value for all of these points to make the comparison. A sufficiently detailed representation of the distribution can be obtained with just a few thousand such estimates, and hence the process was only repeated 2000 times. A "p-value" can be determined by observing what proportion of the T^2 values calculated in the permutation test exceed the observed value.

Additional tests to determine the degree of spatial association between the two samples were performed using variations on the nearest neighbor statistic, as discussed by Sorenson [23]. The simple nearest neighbor statistic is the mean of the distances from each point to the point closest to it. In Sorensen's first approach, originally attributed to Pielou [24], nearest neighbor distances are evaluated by a contingency table. For each point in each group, it is determined whether the nearest neighbor is in the same

group, or in the opposite group. In the present study, this means that for each live birth, it is determined whether the nearest neighbor is also a live birth, or whether it is a birth defect. Similarly, for each birth defect it is determined whether the nearest neighbor is another birth defect or a live birth. A contingency table can then be set up as follows:

		BASE POPULATION	
		Live Births	Birth Defects
NEAREST NEIGHBOR	Live Births	n_1	n_2
	Birth Defects	n_3	n_4

Sorenson and Pielou suggested that the contingency table data could then be evaluated by a χ^2 test. It should be noted that such a test is not statistically correct, since the cells of the contingency table are not independent. In addition, while this method provides some measure of spatial association, the loss of distance information by converting the continuous nearest neighbor distances to a dichotomous situation is undesirable. To avoid this loss of data, Sorensen suggests use of the coefficient of spatial association C_s. In this method, four nearest neighbor distances are calculated: the sum of the distances from a point in group 1 to the nearest neighbor in group 1 (d_{11}), the sum of the distances from a point in group 1 to the nearest neighbor in group 2 (d_{12}), the sum of the distances from a point in group 2 to the nearest neighbor in group 2 (d_{22}), and the sum of the distances from a point in group 2 to the nearest neighbor in group 1 (d_{21}). These four sums correspond to the cells of the contingency table above. The mean distance to a member of the same group is then given by equation 3.2, and the mean distance to a member of the other group is given by equation 3.3.

$$a = \frac{(d_{11} + d_{22})}{N} \tag{3.2}$$

$$b = \frac{(d_{12} + d_{21})}{N} \tag{3.3}$$

The coefficient of spatial association is then given by equation 3.4, where C_s ranges from -1 to 1. Sorenson notes that no statistical test of significance exists for C_s, but that values approaching -1 indicate high dissociation, and values approaching 1 indicate high association [23]. Values near 0 indicate an indeterminate relationship, where each point is equidistant from the nearest neighbor in the same and in the complementary group. The program written to perform both of these nearest neighbor calculations (Nneigh4.c) appears in appendix B.

$$C_s = \frac{a-b}{a+b} \tag{3.4}$$

To test the sensitivity of Hotelling's T^2 to detect variation in the two populations, a series of simulated clusters was introduced to the birth defects population. It is expected that if the methods are sensitive to differences in the two populations, then the introduction of such simulated clusters will result in the statistics indicating that the two populations are not the same, and that some systematic difference exists between the two populations. The clusters were created around fixed points at varying distances from the pooled centroid of the points. All clusters had a radius of 5000 feet (approximately one mile), a distance selected to be consistent with previously reported clusters (e.g. [25]). A total of 20 additional points were added to the map, for each cluster simulation. By using this many points, the clusters are clearly visible in plotted maps of the study area (see figures 3.3-3.11). As such, it would be expected that spatial statistics would detect this clustering as making the modified maps of birth defects dissimilar from the map of live births.

These clusters represent a bias in the spatial distribution of the comparison group that could result if a particular defect or group of defects were to cluster, either as a result of environmental causes such as airborne or hazardous material exposure, or for other reasons.

In this example, only a sample of the total population of birth defects occurring in the county was taken. Since there are few environmental agents known to cause a wide variety of birth defects (so-called "universal teratogens"), to explain the inclusion of these clustered defects one could assume that some form of sampling bias occurred. Either the region where the cluster occurred was sampled more frequently, or some particular defect or group of defects that had clustered was sampled more frequently. Such clustering could also reflect a clustering of mothers at high risk for giving birth to a child with a birth defect due to some non-environmental factor.

3.3. Results

Maps of Santa Clara County showing the spatial distribution of the samples of live births and of birth defects are presented in figures 3.1 and 3.2. Maps reflecting each of the simulated clusters are presented as figures 3.3-3.11.

The random sample of birth defects included 118 births with one defect (54%), 51 with two defects (24%), and 47 with three or more defects (22%). Syndromes were identified in 30 (14%) of the births. The defects included in this random sample are listed in table 3.1. The distribution of maternal race for the sample of birth defects, and for all live births in Santa Clara County during 1983-1985 are shown in table 3.2. The average maternal age for the birth defects sample, and for live births in Santa Clara, were 27.1. Of the children in the birth defects sample, 61.4 percent were male, while among

live births in Santa Clara, 51.1 percent were male.

The resulting distributions of the conversion of the coordinates of the children in the two samples to distance from the pooled centroid are described in table 3.3.

Table 3.1
Defects Present in Random Sample of 216
Birth Defects in Santa Clara County

Defect	BPA	# in Sample	% of Defects
Mobius Syndrome	352	2	0.3
Abnormalities of Jaw	524	9	1.4
Spina Bifida	741	2	0.3
Nervous System Anomalies	742	22	3.4
Anomalies of the Eye	743	35	5.3
Anomalies of the Ear, Face and Neck	744	61	9.3
Bulbus Cordia and Defects of Cardiac Septal Closure	745	40	6.1
Other Cardiac Defects	746	21	3.2
Other Anomalies of Circulatory System	747	29	4.4
Respiratory Defects	748	22	3.4
Cleft Lip/Cleft Palate	749	22	3.4
Other Anomalies of Alimentary Tract	750	63	9.6
Digestive System Defects	751	21	3.2
Anomalies of Genital Organs	752	59	9.0
Anomalies of Urinary System	753	24	3.7
Musculoskeletal Defects	754	73	11.2
Other Anomalies of Limbs	755	41	6.3
Other Anomalies of Skull and Face Bones	756	38	5.8
Anomalies of Integument	757	42	6.4
Chromosomal Anomalies	758	18	2.7
Other Defects	759	5	0.7
Fetal Alcohol Syndrome	760	2	0.3
Congenital Infection	771	4	0.6
Totals		655	100

Table 3.2
Maternal Race in Random Sample of 216
Birth Defects and 201 Live Births
in Santa Clara County, 1983-1985

Race	# of Children in Defects Sample	% of Children in Defects Sample	# of Live Births in Sample.	% of Live Births in Sample
white nonhispanic	176	81.5	117	54.1
white hispanic	8	3.7	38	17.6
black	1	0.5	24	11.1
asian	17	7.9	1	0.5
other	14	6.5	36	16.7

Table 3.3
Univariate Characteristics of
Live Birth and Birth Defect Samples

Live Birth Data	Birth Defect Data
n=201	n=216
$\bar{d}$ = 5.86 miles	$\bar{d}$ = 5.88 miles
$var(\bar{d})$ = 22.33	$var(\bar{d})$ = 17.08
minimum $\bar{d}$ = 0.60 miles	minimum $\bar{d}$ = 0.24 miles
maximum $\bar{d}$ = 28.257 miles	maximum $\bar{d}$ = 27.59 miles

The Kolmogorov-Smirnov statistic is the maximum absolute difference between the two cumulative distributions. Calculation of the Kolmogorov-Smirnov statistic was performed using the SPSSX computer program [21]. The observed value for the Kolmogorov-Smirnov statistic was 0.0745. The z-value associated with this statistic was 0.76 with a 2-tailed p-value of 0.610. The cumulative distributions of the live birth and birth defect samples are shown in figure 3.12.

The analysis of distances with a univariate t-test resulted in a t value of -0.03, with an associated p-value of 0.978.

The observed Hotelling's T^2 for the samples of 201 live births and 216 birth defects was 0.286, with a p-value of 0.867, when converted according to equation 3.1 and compared to an F distribution with 2 and 415 degrees of freedom. The permutation approach yielded a distribution of T^2 values with

a mean of 1.97 and a standard deviation of 1.99. The randomized p-value, that is, the probability that a T^2 exceeds the observed value of 0.286 based on the distribution of the permutation, is 0.865.

The contingency table approach to nearest neighbor distances resulted in the following table:

		BASE POPULATION	
		Live Births	Birth Defects
NEAREST NEIGHBOR	Live Births	86	103
	Birth Defects	115	113

The χ^2 for this table is 1.008, with a p-value of 0.685. The coefficient of spatial association, C_s was calculated to be 0.044.

The results of the sensitivity tests are next presented. Univariate analysis by using the Kolmogorov-Smirnov test on each of the nine clusters is presented in Table 3.4. The distance from the cluster to the pooled centroid of the birth defects and live births in the initial analysis is presented. The comparison group in all cases was the unmodified sample of live births. These results are presented pictorally in figures 3.13-3.21.

Table 3.4
Kolmogorov-Smirnov Analysis of Modified Birth Defects Maps

Dataset	Distance from cluster to centroid (miles)	Kolmogorov Smirnov Statistic	K-S Z	p-value
Live Births	--	--	--	--
Birth Defects	--	0.075	0.76	0.610
First Cluster	0	0.105	1.096	0.181
Second Cluster	1	0.093	0.969	0.304
Third Cluster	2	0.078	0.816	0.519
Fourth Cluster	3	0.045	0.469	0.980
Fifth Cluster	4	0.097	1.008	0.261
Sixth Cluster	5	0.107	1.116	0.166
Seventh Cluster	10	0.109	1.132	0.154
Eighth Cluster	15	0.109	1.132	0.154
Ninth Cluster	20	0.109	1.132	0.154

The results of the T^2 test comparing the cluster investigations established in the birth defects group to the live birth sample are presented in table 3.5. The distance of each cluster to the pooled centroid of both groups is shown, along with the centroid of the group on the map scale (1 map unit=1000 feet).

Table 3.5
Multivariate Analysis of Santa Clara Map

Dataset	Distance from cluster to centroid (miles)	$\bar{x}$ (map units)	$\bar{y}$ (map units)	Hotelling's T^2	p-value
Live Births	--	93.977	300.189	--	--
Birth Defects	--	95.433	299.483	0.286	0.87
First Cluster	0	94.719	299.797	0.277	0.87
Second Cluster	1	98.438	296.062	0.425	0.81
Third Cluster	2	102.144	292.301	0.611	0.74
Fourth Cluster	3	105.849	288.540	0.834	0.66
Fifth Cluster	4	109.555	284.778	1.092	0.58
Sixth Cluster	5	113.261	281.017	1.383	0.50
Seventh Cluster	10	131.789	262.211	3.230	0.20
Eighth Cluster	15	150.317	243.405	5.432	0.07
Ninth Cluster	20	168.845	224.599	7.638	0.02

3.4. Discussion

Two major points become clear from the results of these analyses. First, there is little evidence to support the alternative hypothesis that the population of live births and the population of birth defects in Santa Clara County are spatially different, in some systematic fashion. Second, the spatial statistics available for use in this study are generally insensitive to detect these specific perturbations in the spatial pattern of birth defects, as evidenced by their failure to detect most of the simulated clusters introduced to the maps—a clustering relatively obvious by simple examination of the maps.

One source of potential bias in the methodology used in this study is the process of digitization. While the computer hardware for this process is accurate, the identification of the exact location of a street address using assessors maps and commercial street maps is not precise. There is reasonable confidence that digitized locations are near the actual location of the address digitized (within 500 feet). Due to imprecise street number notation

on the assessor's maps, it is possible that some addresses were shifted one or two houses from their true location. On the scale used for this analysis, this error is minor. In addition, since the error introduced is random, and as it applies equally to both the live birth locations and to the locations of children born with a birth defect, it is not thought that this error compromises the results of this study.

Another potential source of bias related to the address data used is associated with the difference in address sources. The addresses used for live births were those recorded on the birth certificate; information collected from the parents at the time of delivery. Addresses used for children with a birth defect were abstracted from the medical record, and may not be the same as that shown on the birth certificate. No comparison has yet been made between addresses in the CBDMP's registry and those on the birth certificate. Such a comparison would help clarify whether any appreciable difference exists between the two sources. As the bias introduced here was random, with respect to spatial distribution, it was unlikely to be a significant problem in this analysis. A possibly more serious problem associated with address is the migration between first trimester and birth, as discussed in chapter 1. If this migration is differential with respect to birth outcome, the resulting measures of association would be biased. The effect of migration was not assessed in the present study.

None of the statistics used yielded strong evidence for differences between the sample of live births and the unmodified sample of all birth defects. No major differences can be seen in the maps showing the two samples (figures 3.1 and 3.2). The mean difference in distance to the pooled centroid was approximately 20 feet, which is less than one house-length. While the sample of birth defects has a slightly smaller variance

compared to that of all live births, the difference in the standard deviations is small compared to the scale of the county, and thus this variation is probably unimportant. Indeed, a t-test comparing the distributions of the distances to the pooled centroid yielded a p-value of 0.978. The Kolmogorov-Smirnov test also yielded results consistent with no measurable difference between the two maps, with a p-value of 0.67. Nearest neighbor methods yielded corresponding results, and although no p-value can be attached to the coefficient of spatial variation, C_s, the observed value of 0.044 is close to 0, which is consistent with no association or dissociation. The variability of the p-values reflects the different measures being used, and is overshadowed by the lack of significance of all values.

The Hotelling's T^2 test gave a p-value of 0.867. As this can be considered to be a refined measure of the univariate t-test, this p-value is consistent with the value of 0.978 noted above. The simulated p-value generated using the permutation test of 0.865 is close to the p-value from the F-distribution, indicating that the T^2 statistic based on the assumption of normally distributed data is robust and simulated p-values are not necessary.

The results of the sensitivity analysis present a discouraging view of the ability to determine whether two populations are interchangeable. The introduction of highly significant clusters was usually not detected by either the Kolmogorov-Smirnov test, or the T^2 test. In both cases, the distance between the cluster and pooled centroid affected both the value of the statistic, and the statistic's ability to detect the change. In the case of the Kolmogorov-Smirnov test, when the cluster occurred within two miles of the centroid, the cumulative distribution of birth defects can be seen to rise rapidly above that of live births at the point of the cluster, providing the maximum absolute difference used in this method (figures 3.13-3.15).

When the cluster was between three and five miles of the pooled centroid, however, the effect was to make the cumulative distribution of distances of birth defects more closely match the distribution of distances of live births (figures 3.16-3.18). This is because in the initial map, the cumulative distribution of birth defects is below that of live births for distances between approximately three and six miles, and hence adding 20 birth defects shifts the cumulative birth defects distribution, making it roughly collinear with that of the live births.

When the cluster occurs at greater than 10 miles from the pooled centroid (figures 3.19-3.21), the maximum absolute difference in the two cumulative distributions is caused by the cluster, as can be seen by the sudden increase in birth defects at the distance of the cluster from the centroid in the figures. All three absolute differences happen to be the same, and the statistic is thus constant for the last 3 clusters.

The Hotelling's T^2 test is also sensitive to the location of the cluster in relation to the pooled centroid. This is because this test functions in the same way as a univariate t-test, measuring the distance between the means (in two dimensions, the distance between the centroids). In this demonstration, a fixed number (20) of additional birth defect children were added to each modified map. As such, the further these clustered points are from the pooled centroid, the more the centroid of all 236 birth defects will be moved from its initial value. Indeed, for this particular situation, the cluster must be approximately 17 miles from the pooled centroid to be detected at a level of α=0.05. The change in T^2 as distance increases is shown in figure 3.22.

The lack of sensitivity of these methods and the sensitivity to distance of the perturbation from the centroid weaken the argument that live births and birth defects are interchangeable as comparison groups for the

investigation of clusters, but this possibility cannot be ruled out. Certainly no dramatic difference in the two spatial distributions exists, and further investigation using the spatial distributions of the two populations interchangeably is not ruled out. At this stage, evidence for discounting further investigation using the two populations interchangeably might include visual evidence that the two populations are spatially concentrated in different parts of the county, or differences between the populations as detected by one of the methods used in this chapter. While more sensitive methods are necessary to determine the extent to which these two populations may be different, the development of such methodology is beyond the scope of this presentation.

The use of a surrogate population in the spatial investigation of clusters would be a beneficial tool to health departments, registries, and others who are charged with responding rapidly to inquiries from the public about whether clustering is going on. Since spatial data on surrogate populations such as all birth defects are often readily available in electronic form in the same registry where the "numerator" data are contained, expensive and time-consuming additional data collection could be avoided, and a rapid response made. In addition, studies previously not possible due to lack of referent data could be performed. This method can also serve as a filter, permitting the rapid separation of investigations which are important to pursue, from an epidemiologic perspective, from those which are less likely to yield meaningful results.

An additional advantage of using surrogates from the same registry stems from the precision of street address. Many studies of clusters have used study areas considerably larger than the actual exposure as data, particularly data on controls, are available in aggregated form at the census tract, zip code or county level. When street address is available, it becomes

possible to more narrowly define the region of exposure, and to identify which cases and which non-cases are actually exposed. The result is a reduction in misclassification, as exposure no longer needs to be assumed to occur over the entire region, which will increase the power of such a study to detect an effect, should one exist. The results presented in this chapter suggests that the use of all birth defects as a surrogate for all live births is reasonable, as the two populations appear to be spatially distributed in a similar manner.

3.5. References

1. Freeman, A. C.; Zucker, R. S. An investigation into a reported cancer cluster using the nearest-neighbor statistical technique. *Prog Clin Biol Res.* 216: 53-58; 1986.
2. Macera, C. A.; Garrison, C. Z.; Mayberry, R. M. Investigation of a cancer cluster in a rural South Carolina area. *J SC Med Assoc.* 1987: 107-110.
3. Tobler, W. R. A continuous transformation useful for districting. *Ann NY Acad Sci.* 219: 215-220; 1983.
4. Selvin, S.; Merrill, D.; Schulman, J.; Sacks, S.; Bedell, L.; Wong, L. Transformations of maps to investigate clusters of disease. *Soc Sci Med.* 26: 215-221; 1988.
5. Berkson, J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bull.* 2: 47-53; 1946.
6. Norell, S. E.; Ahlbom, A. Hospital versus population referents in two case-referent studies. *Scand J Work Environ Health.* 13: 62-66; 1987.
7. Smith, A. H.; Pearce, N. E.; Callas, P. W. Cancer case-control studies with other cancers as controls. *Int J Epidemiol.* 17: 298-306; 1988.
8. Harris, J. A. California Birth Defects Monitoring Program. *Western J Med.* 145: 265; 1986.
9. Holtzman, N. A.; Khoury, M. J. Monitoring for congenital malformations. *Ann Rev Public Health.* 7: 237-266; 1986.
10. United States Bureau of the Census. County and City Data Book. Washington: United States Government Printing Office; 1988.

11. State of California Department of Health Services. Vital Statistics of California 1985. Sacramento, CA: Health and Welfare Agency; 1987.

12. California Birth Defects Monitoring Program. Unpublished Data. 1990.

13. State of California Department of Health Services. Vital Statistics of California 1983. Sacramento, CA: Health and Welfare Agency; 1986.

14. State of California Department of Health Services. Vital Statistics of California 1984. Sacramento, CA: Health and Welfare Agency; 1986.

15. County Profiles, Sources, and General Information. Sacramento: California Department of Commerce; 1986.

16. Census Tract Street Index for Santa Clara County. San Jose, CA: Santa Clara County Planning Office; 1980.

17. Assessor's Maps for Santa Clara County. San Jose, CA: Santa Clara County, Assessor's Office; 1986.

18. Santa Clara County Popular Street Guide, Census Tract Edition. San Francisco, CA: Thomas Brothers, Inc.; 1983.

19. Sigma-Scan Computer Package. Corte Madera, CA: Jandel Software; 1986.

20. Public Resources Code of the State of California. In: *Deering's California Codes.* San Francisco, CA: Bancroft-Whitney; 1987.

21. SPSSX User's Guide. SPSS, Inc.: McGraw Hill; 1983.

22. Johnson, R. A.; Wichern, D. W. Applied Multivariate Statistical Analysis. Englewood Cliffs, NJ: Prentice-Hall; 1982.

23. Sorensen, A. D. A review of techniques for measuring the degree of spatial association between point sets. *New England Research Series in Applied Geography.* 41; 1976.

24. Pielou, E. C. Segregation and symmetry in two-species populations as studied by nearest-neighbor relationships. *J Ecology*. 49: 255-269; 1961.

25. Epidemiological Studies Section. Cardiac Defects in Relation to Water Contamination, 1981-1982, Santa Clara County, California. Berkeley, CA: California Department of Health Services; 1985.

Figure 3.1
Random Sample of 201 Live Births
in Santa Clara County

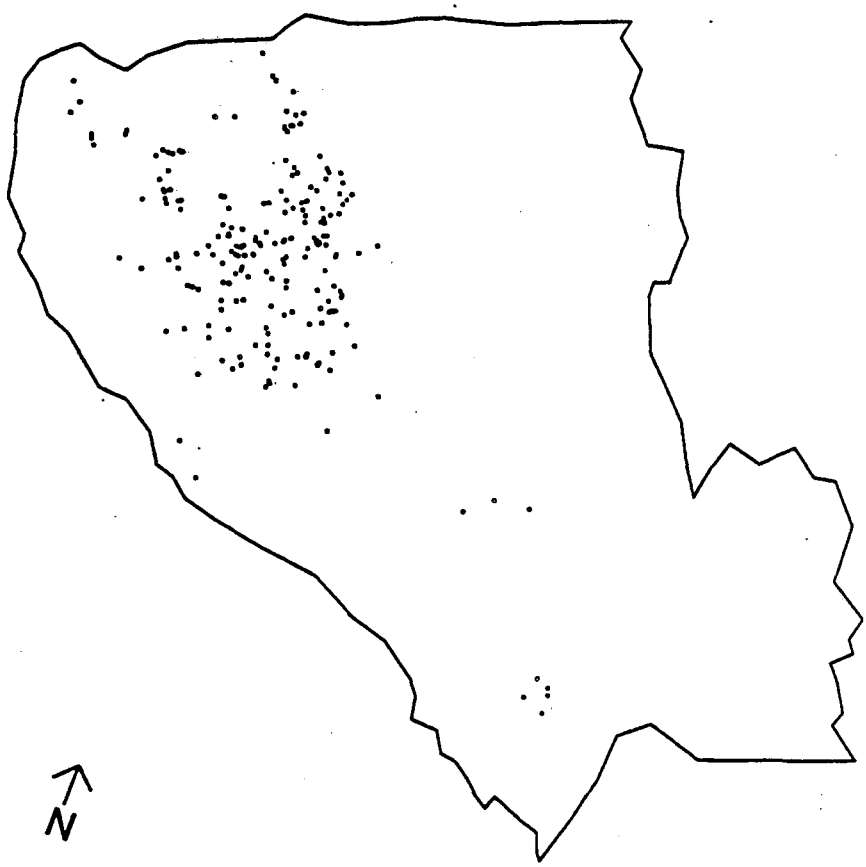


Figure 3.2
Random Sample of 216 Children with Birth Defects in Santa Clara County

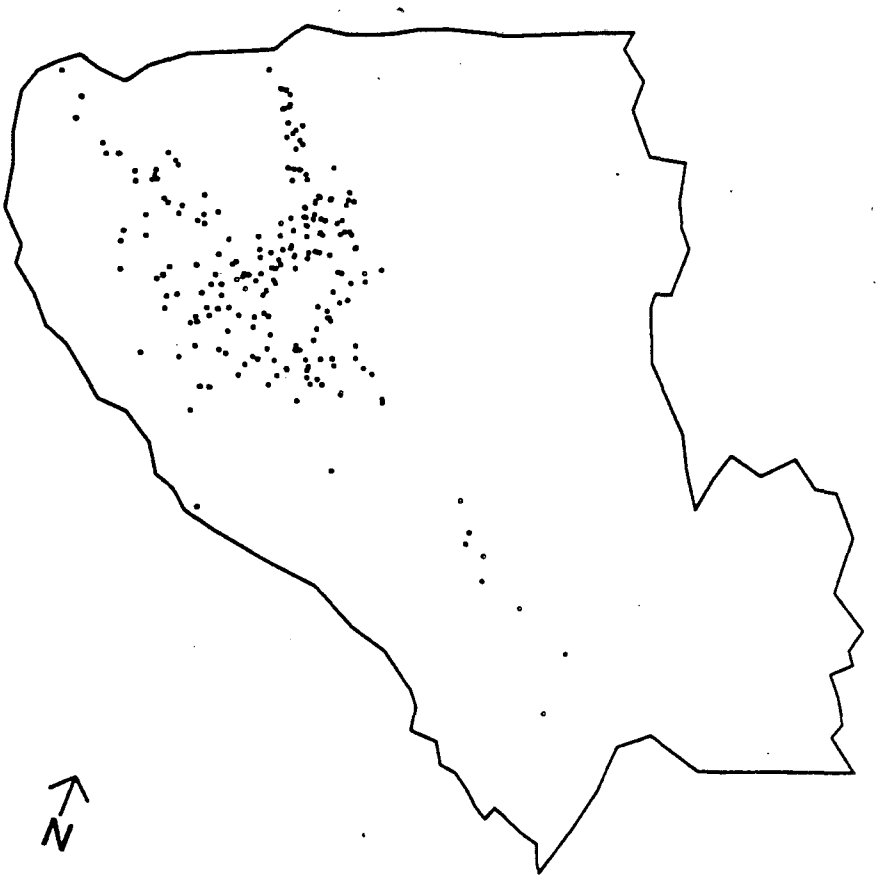


Figure 3.3
Simulated Cluster of 20 Birth Defects
Distance=0 miles from Centroid

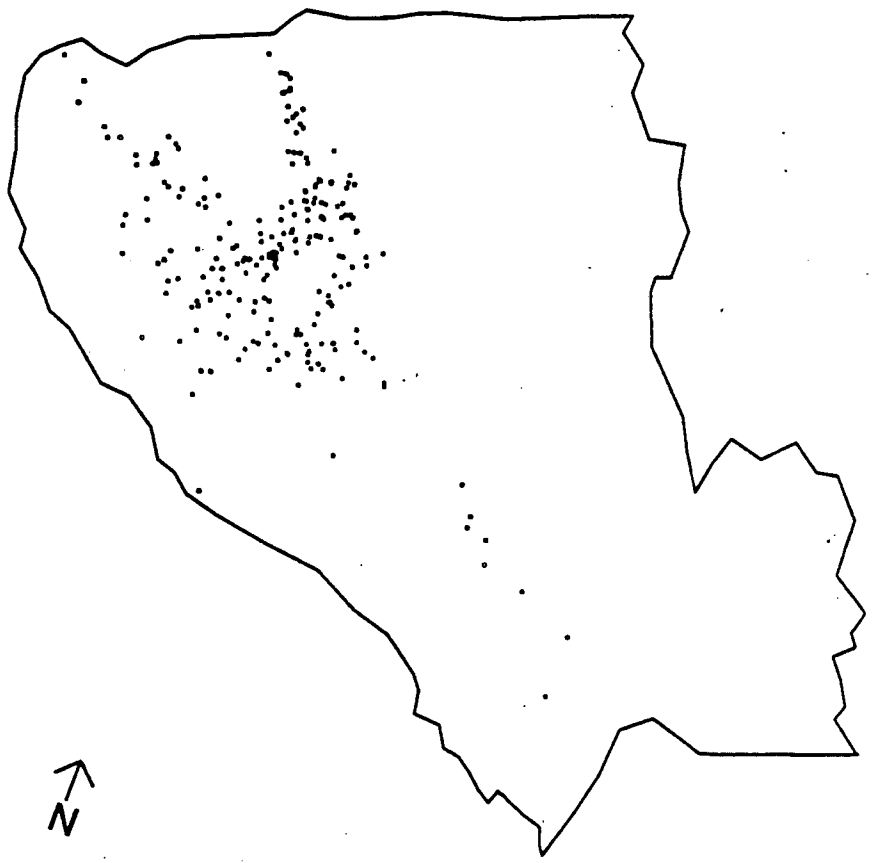


Figure 3.4
Simulated Cluster of 20 Birth Defects
Distance=1 mile from Centroid

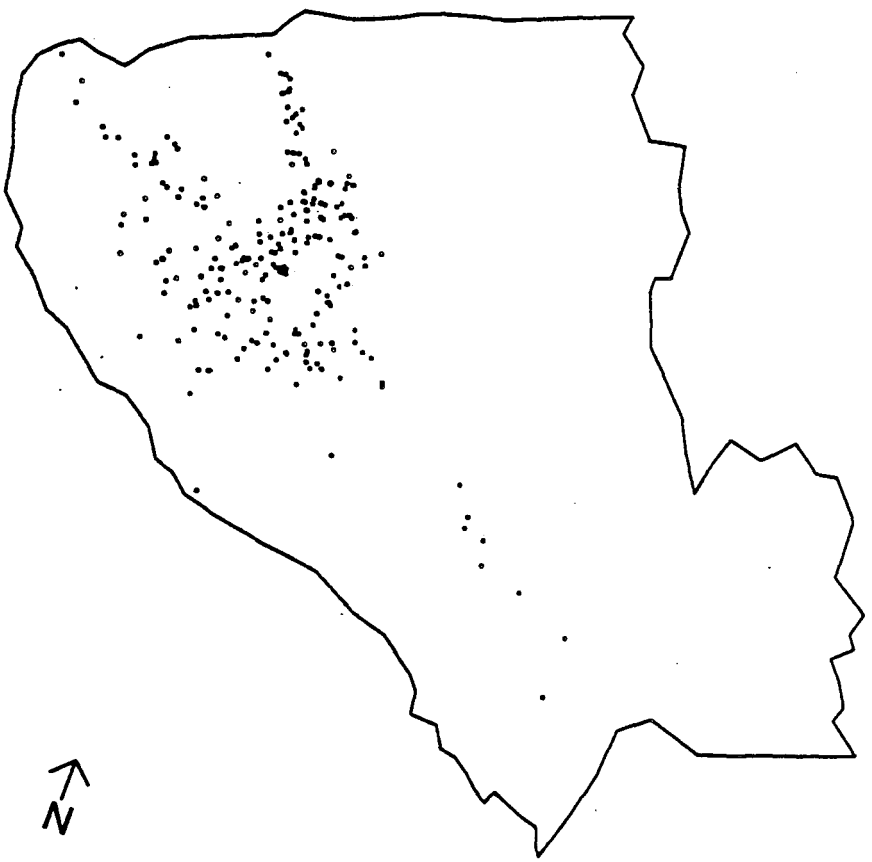


Figure 3.5
Simulated Cluster of 20 Birth Defects
Distance=2 miles from Centroid

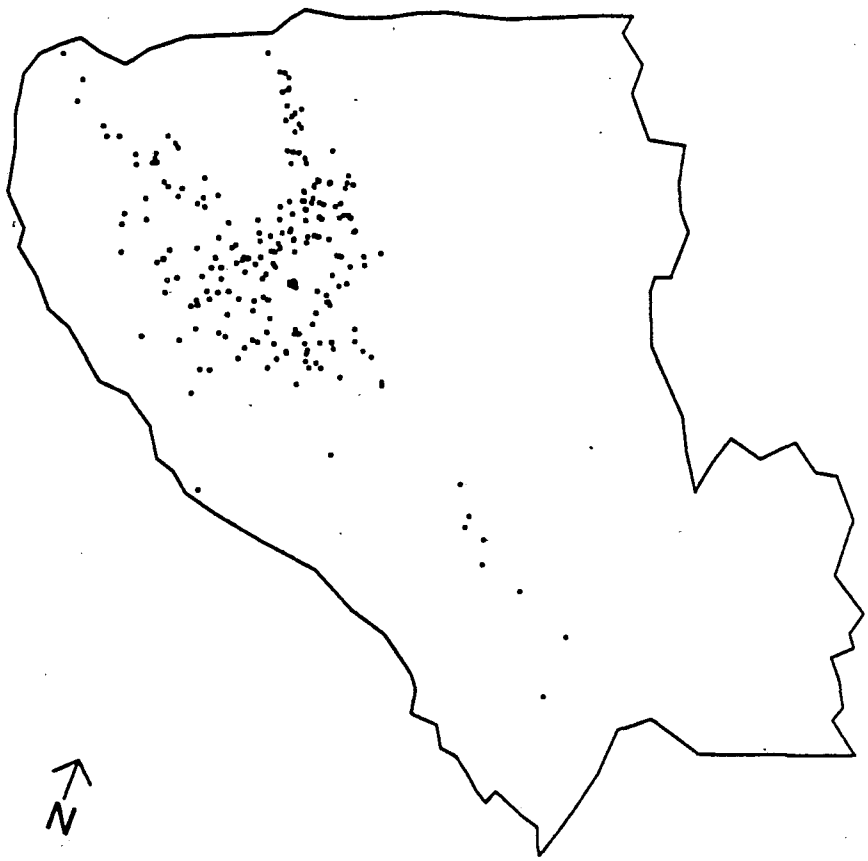


Figure 3.6
Simulated Cluster of 20 Birth Defects
Distance=3 miles from Centroid

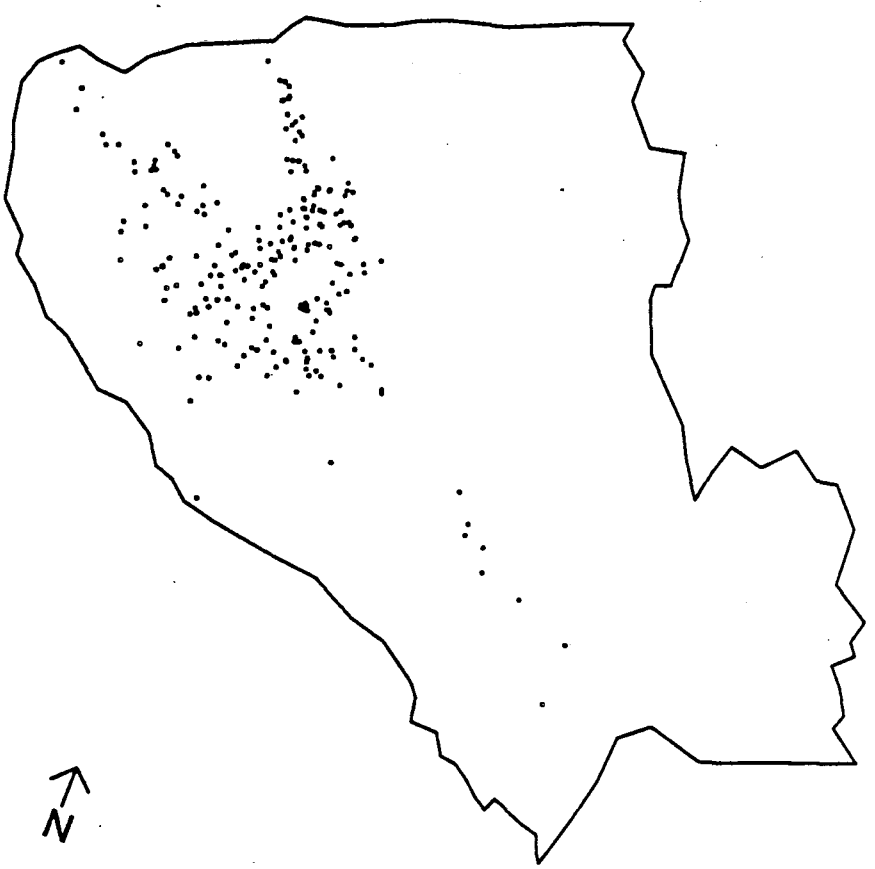


Figure 3.7
Simulated Cluster of 20 Birth Defects
Distance=4 miles from Centroid

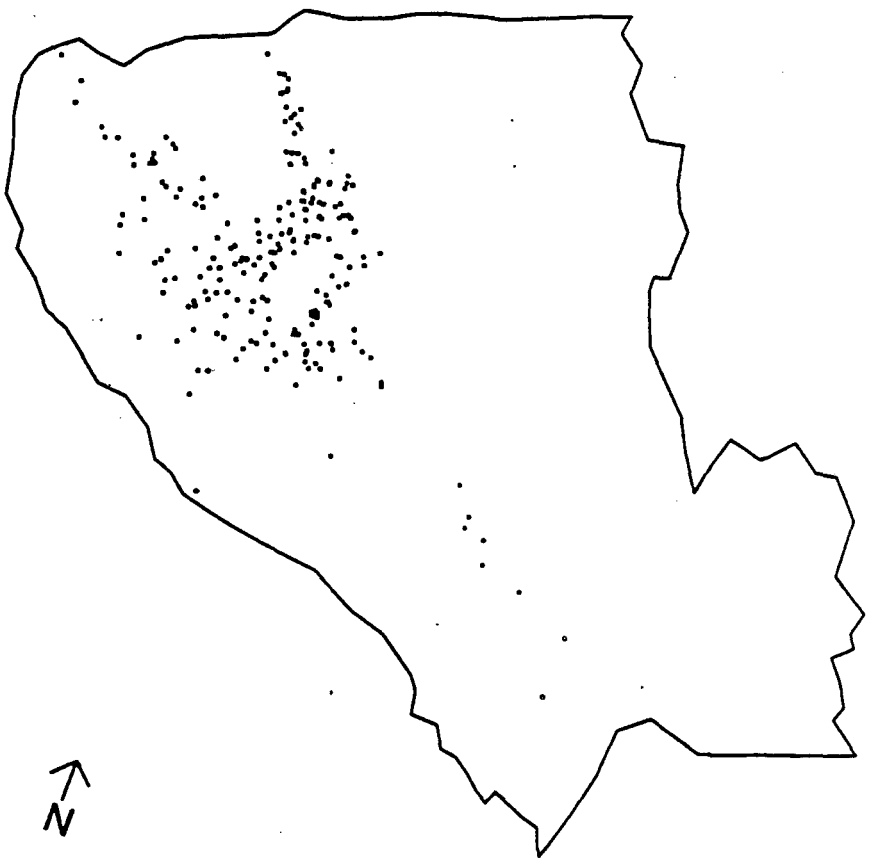


Figure 3.8
Simulated Cluster of 20 Birth Defects
Distance=5 miles from Centroid

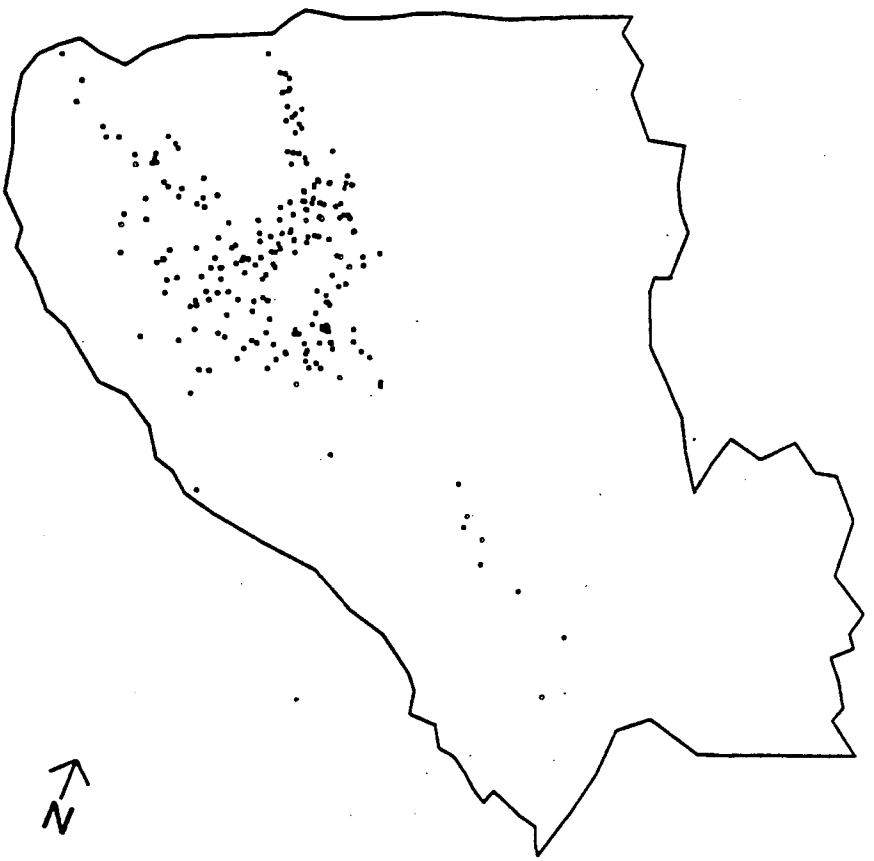


Figure 3.9
Simulated Cluster of 20 Birth Defects
Distance=10 miles from Centroid

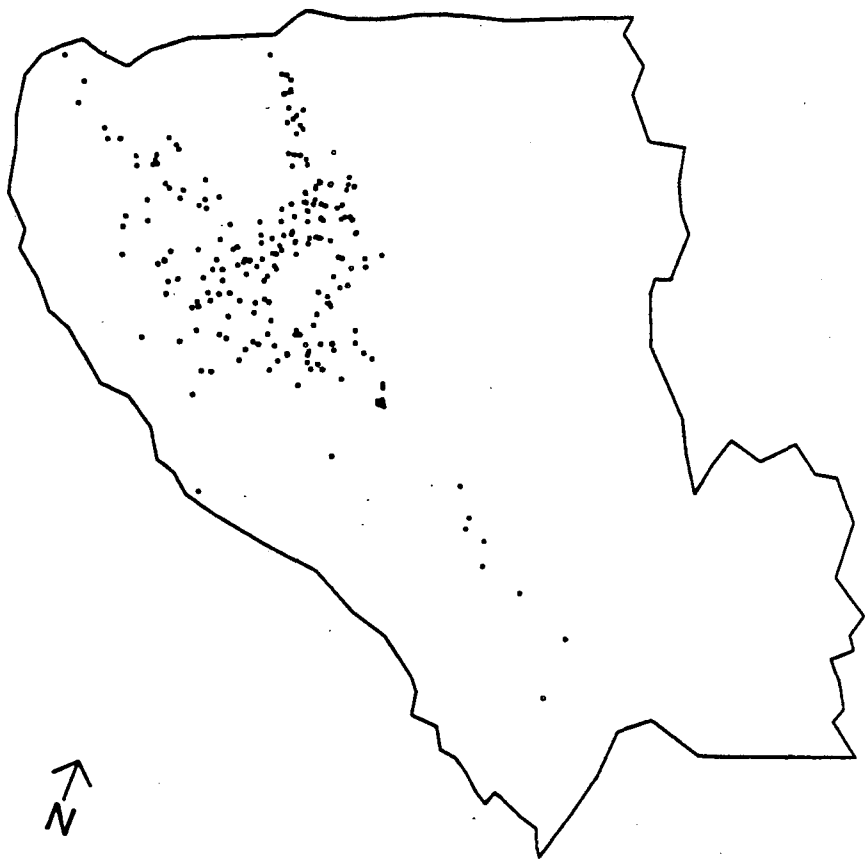


Figure 3.10
Simulated Cluster of 20 Birth Defects
Distance=15 miles from Centroid

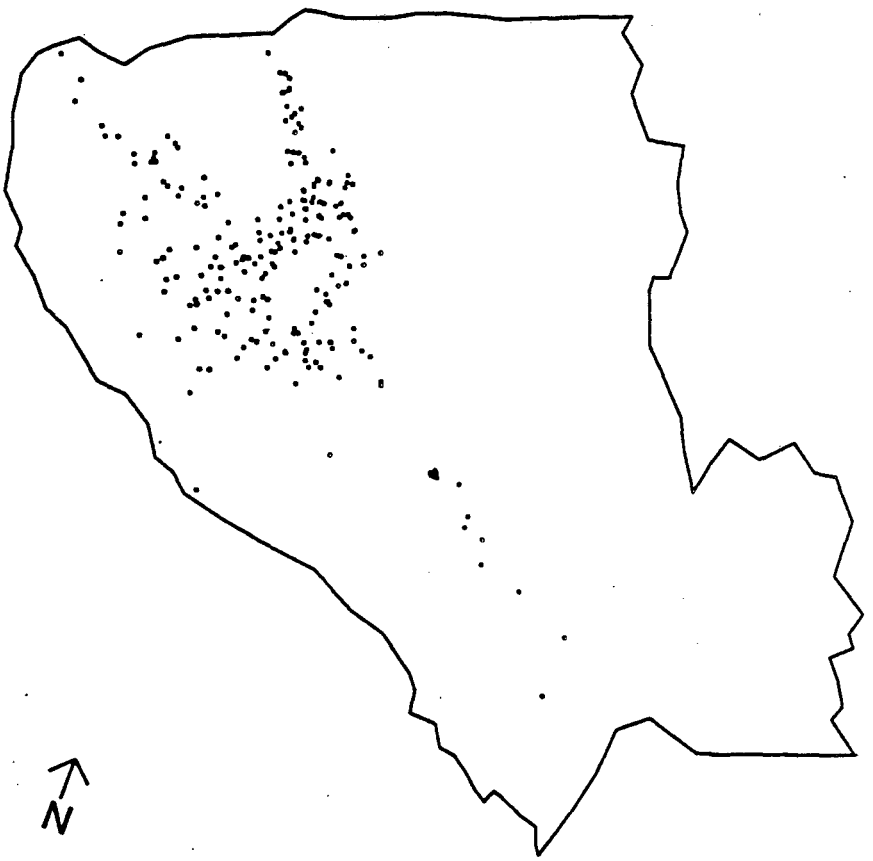


Figure 3.11
Simulated Cluster of 20 Birth Defects
Distance=20 miles from Centroid

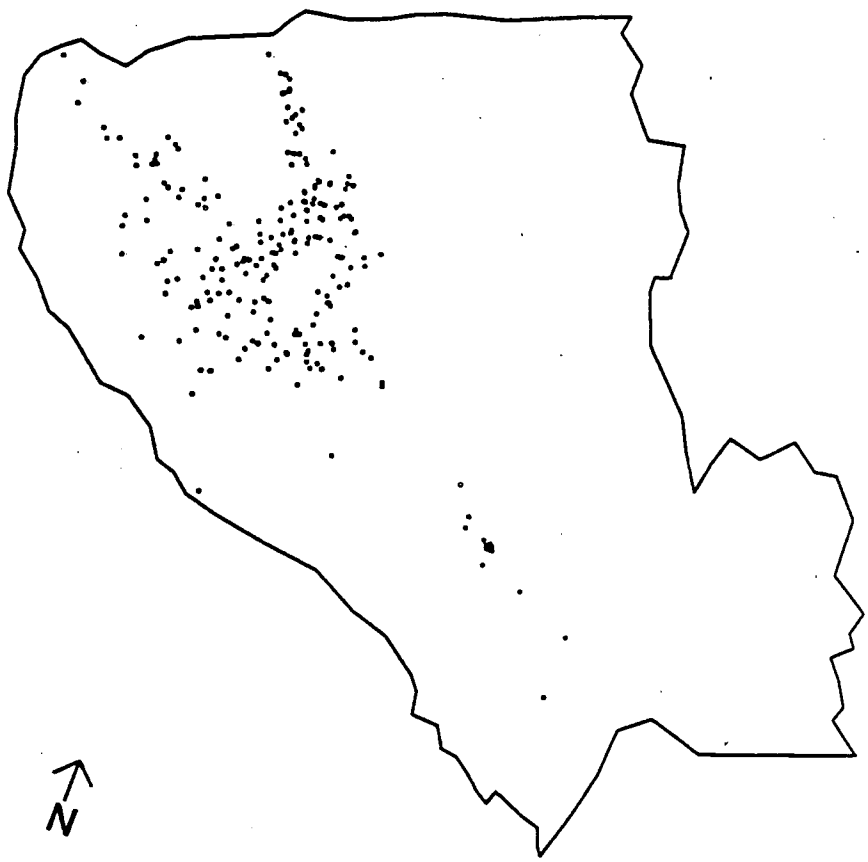


Figure 3.12
Cumulative Probability Distributions
No Cluster

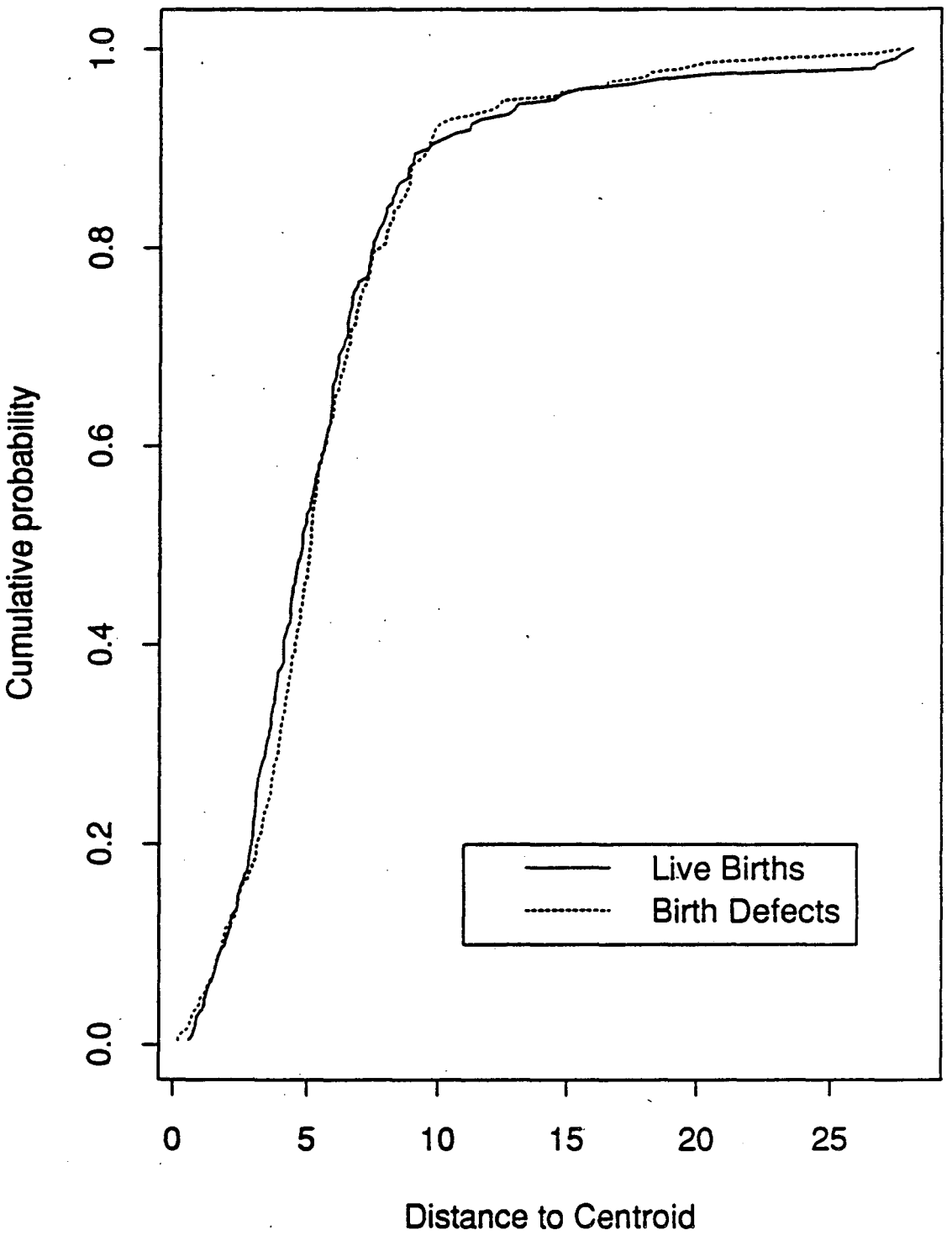


Figure 3.13
Cumulative Probability Distributions
Cluster at D=0 miles

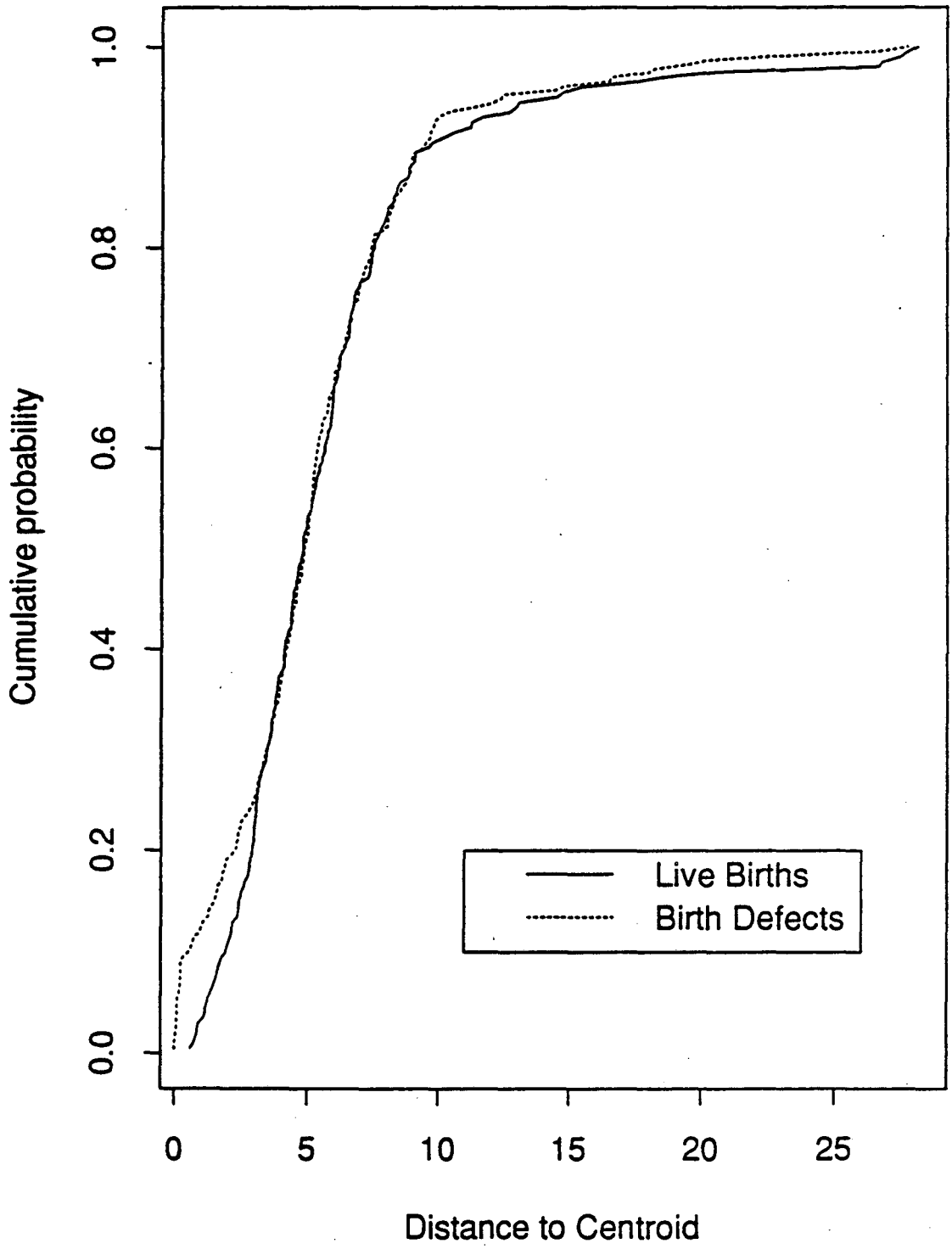


Figure 3.14
Cumulative Probability Distributions
Cluster at D=1 mile

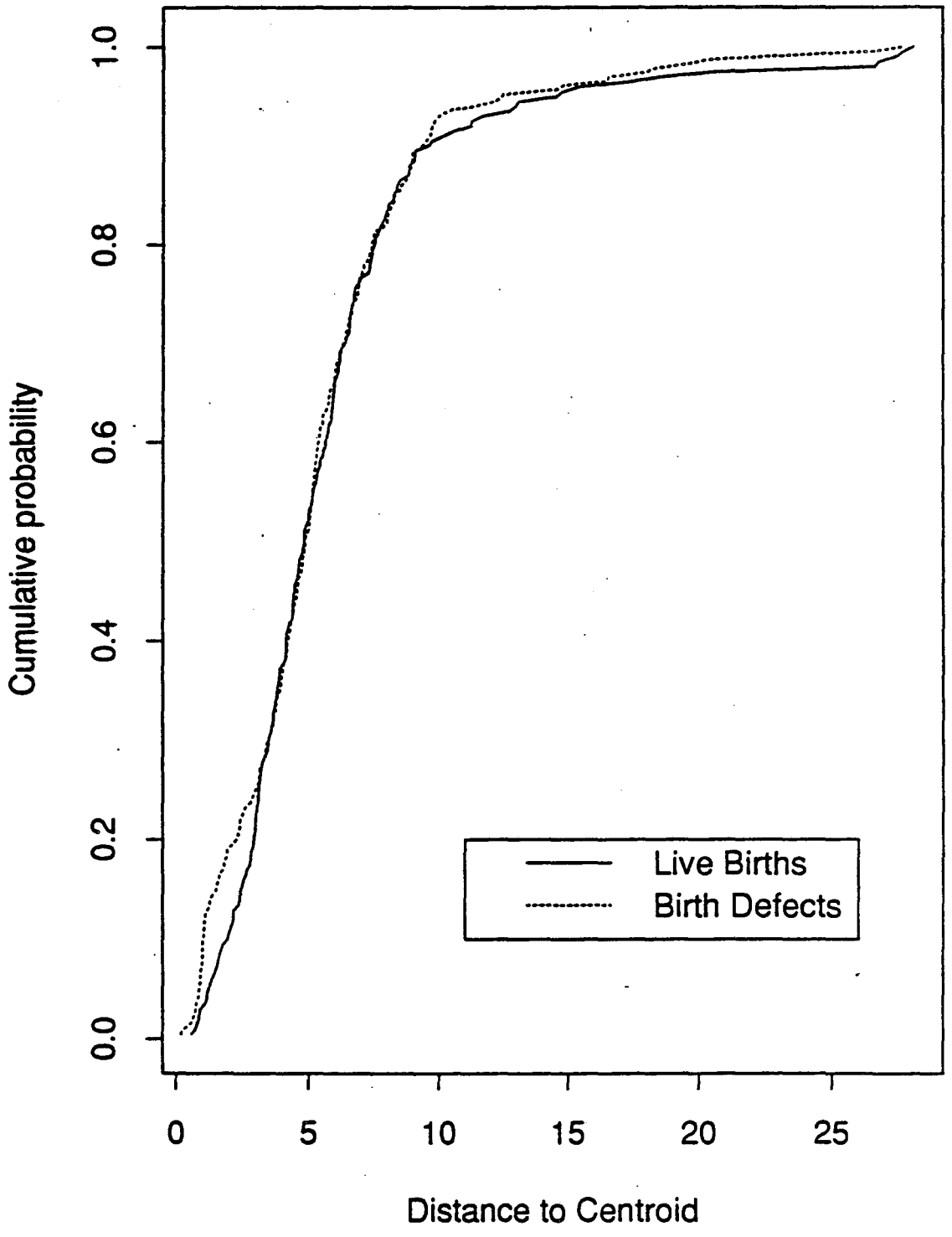


Figure 3.15
Cumulative Probability Distributions
Cluster at D=2 miles

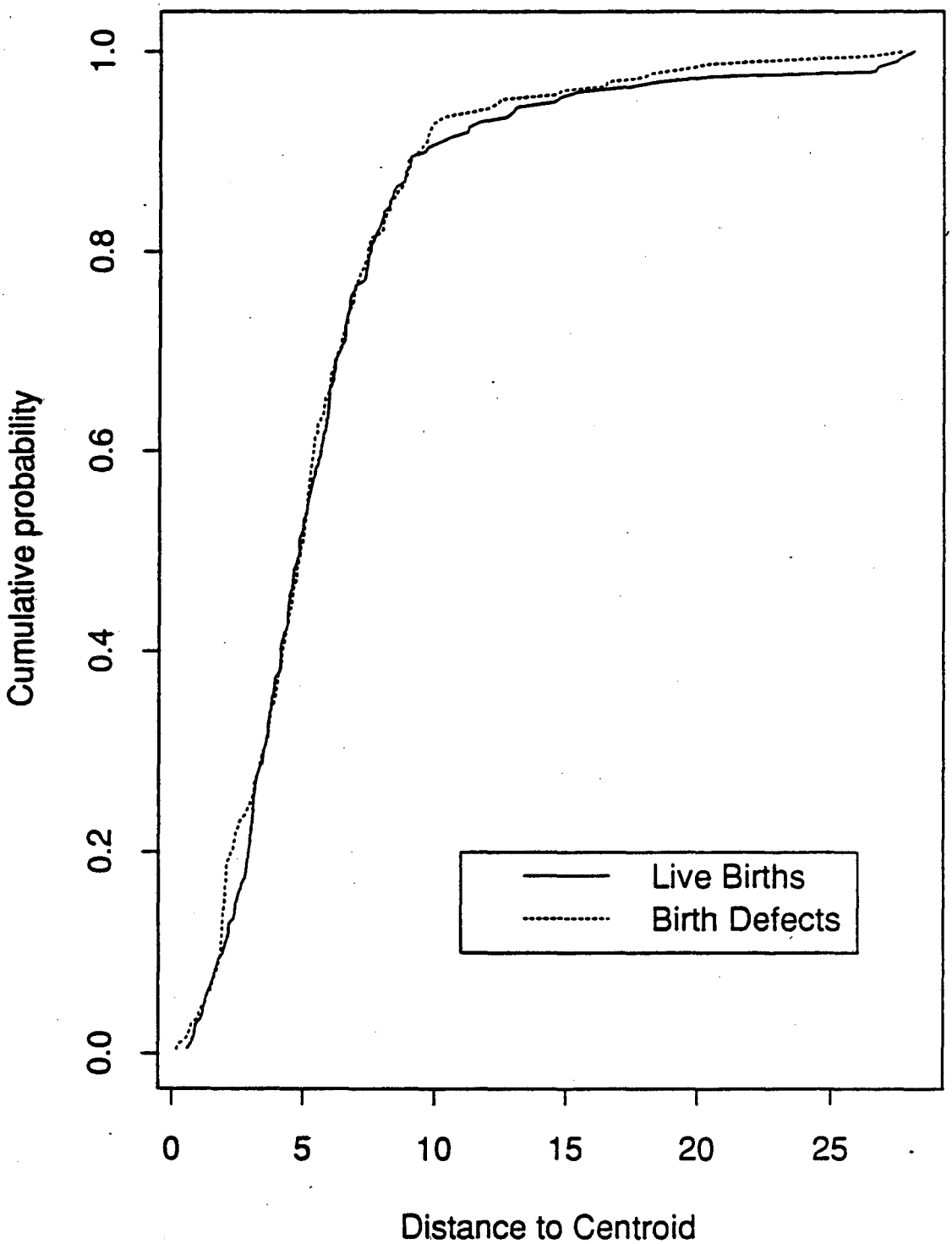


Figure 3.16
Cumulative Probability Distributions
Cluster at D=3 miles

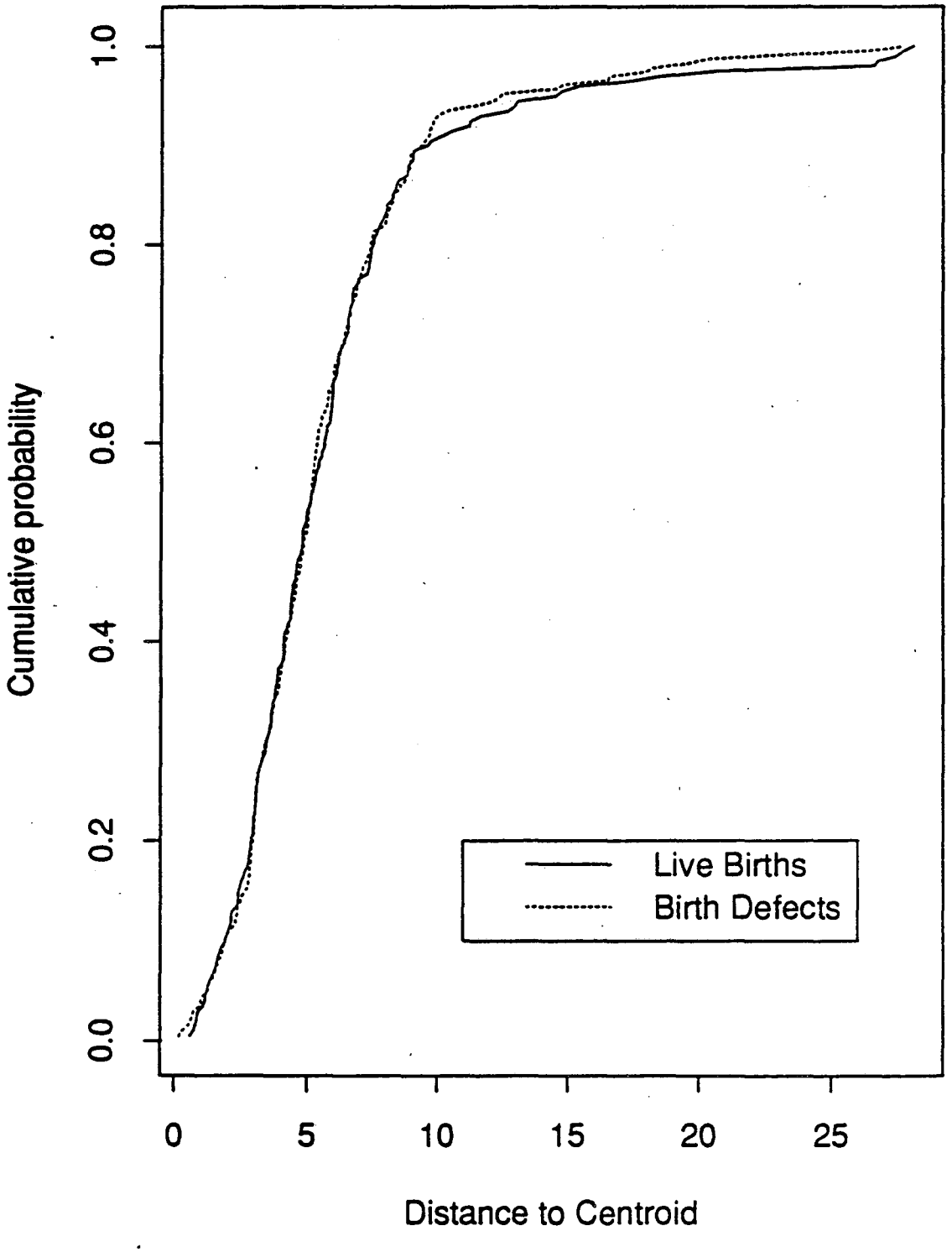


Figure 3.17
Cumulative Probability Distributions
Cluster at D=4 miles

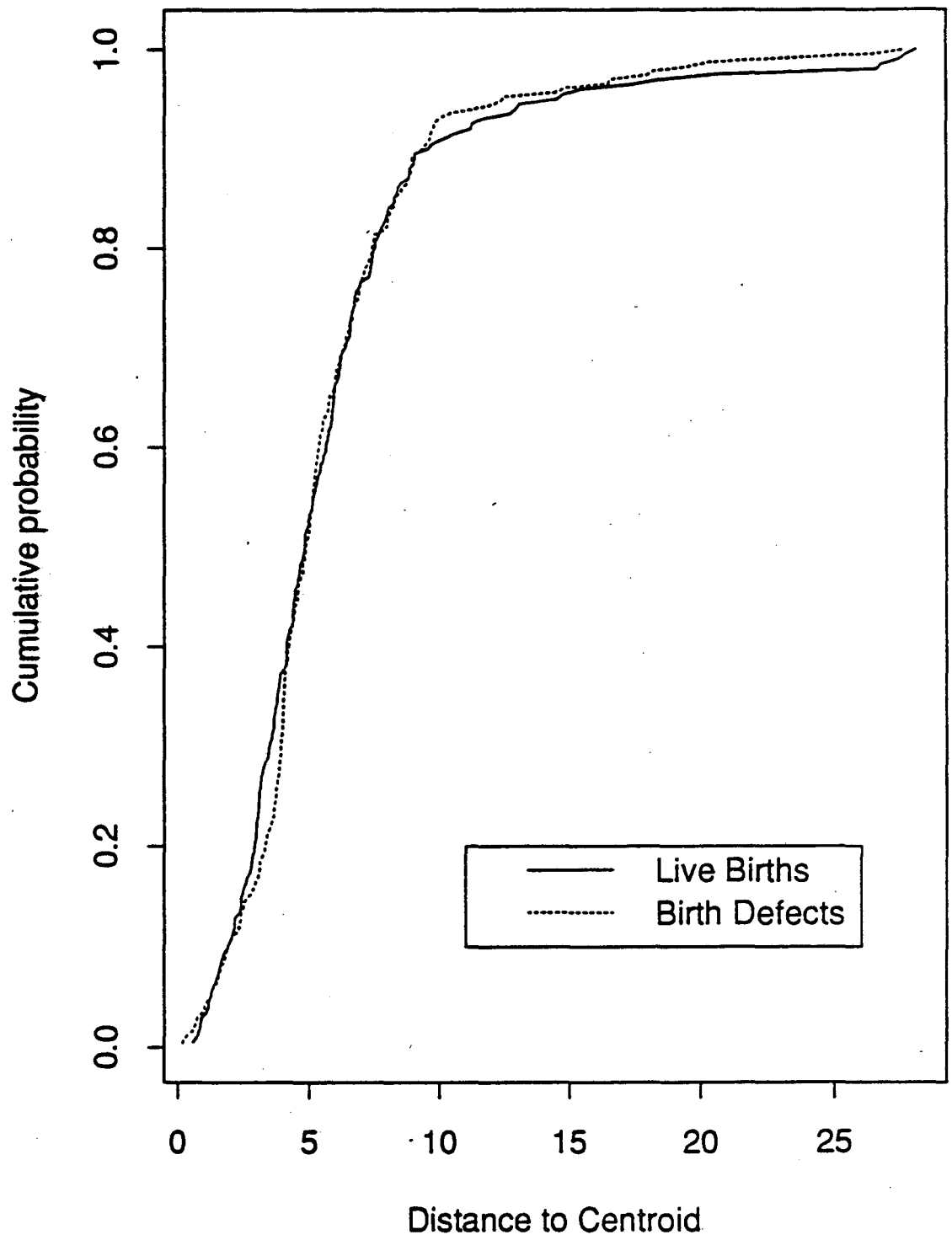


Figure 3.18
Cumulative Probability Distributions
Cluster at D=5 miles

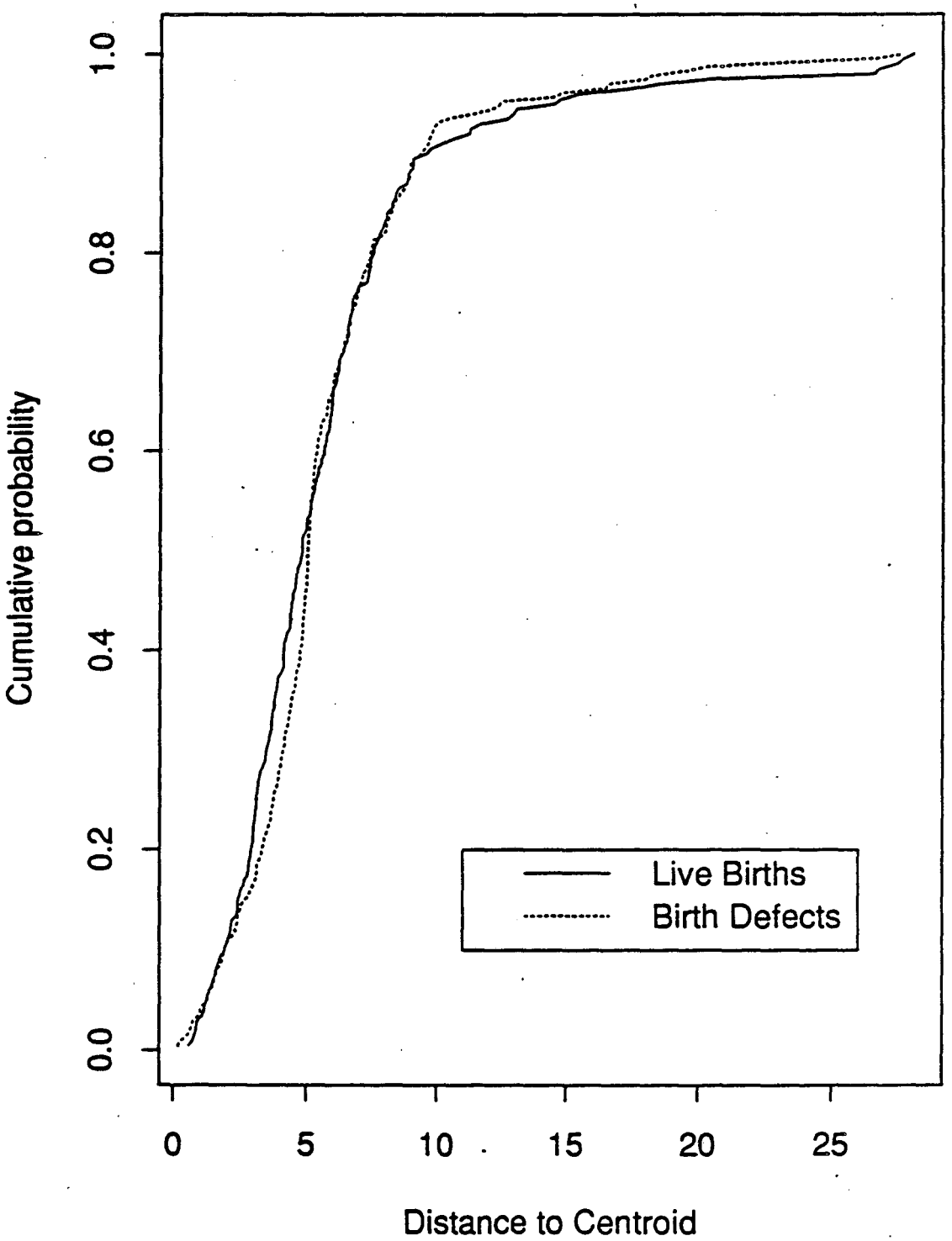


Figure 3.19
Cumulative Probability Distributions
Cluster at D=10 miles

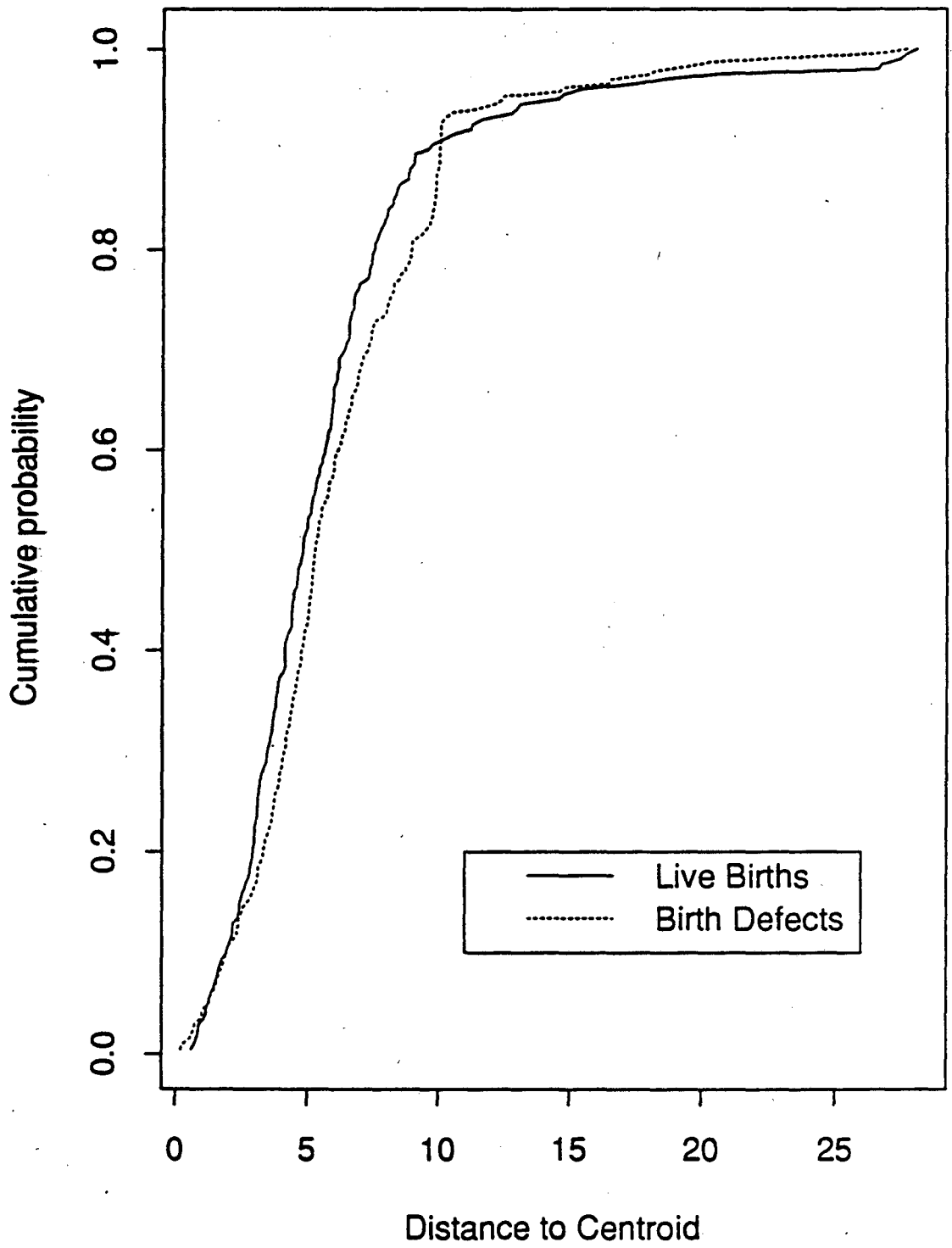


Figure 3.20
Cumulative Probability Distributions
Cluster at D=15 miles

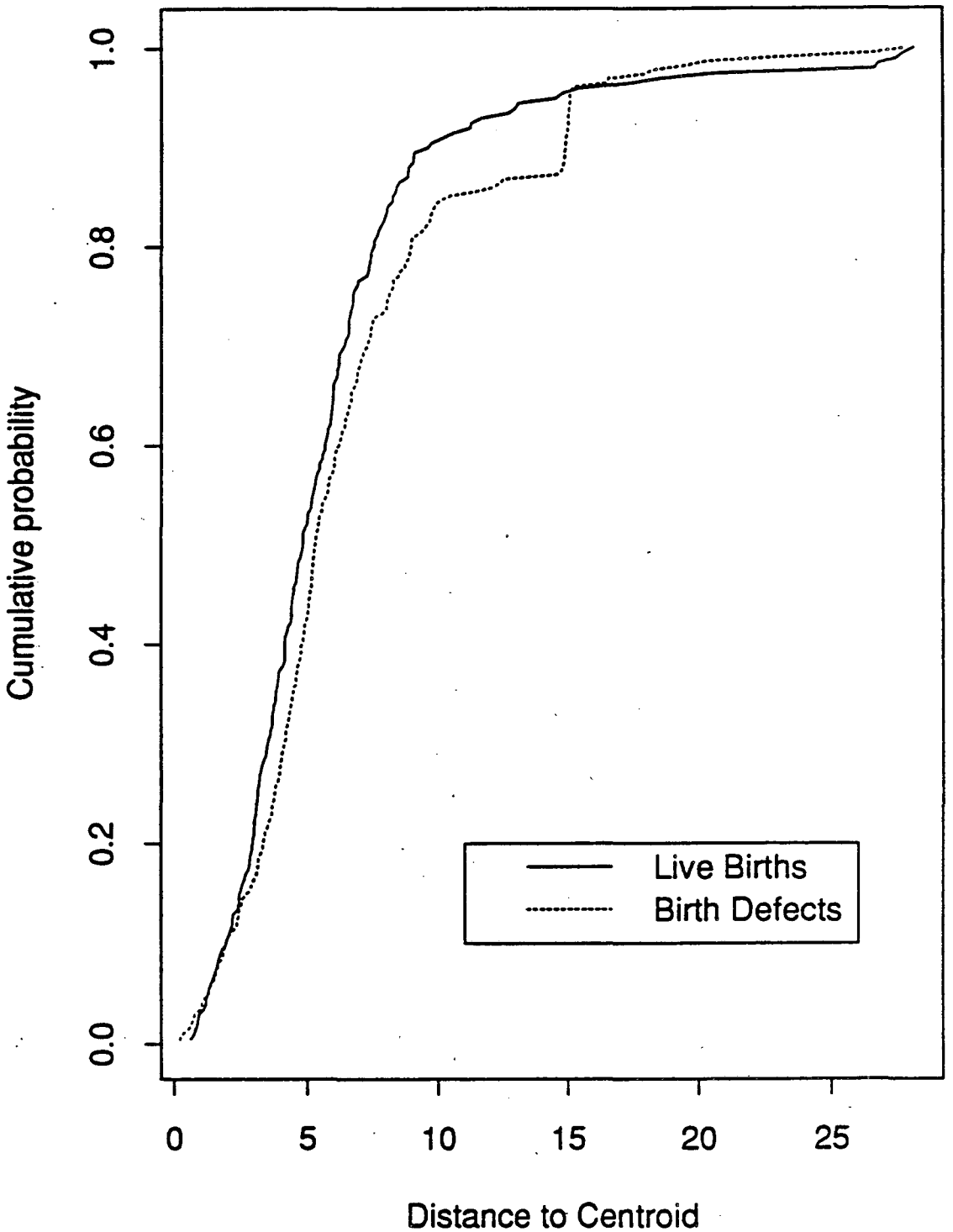
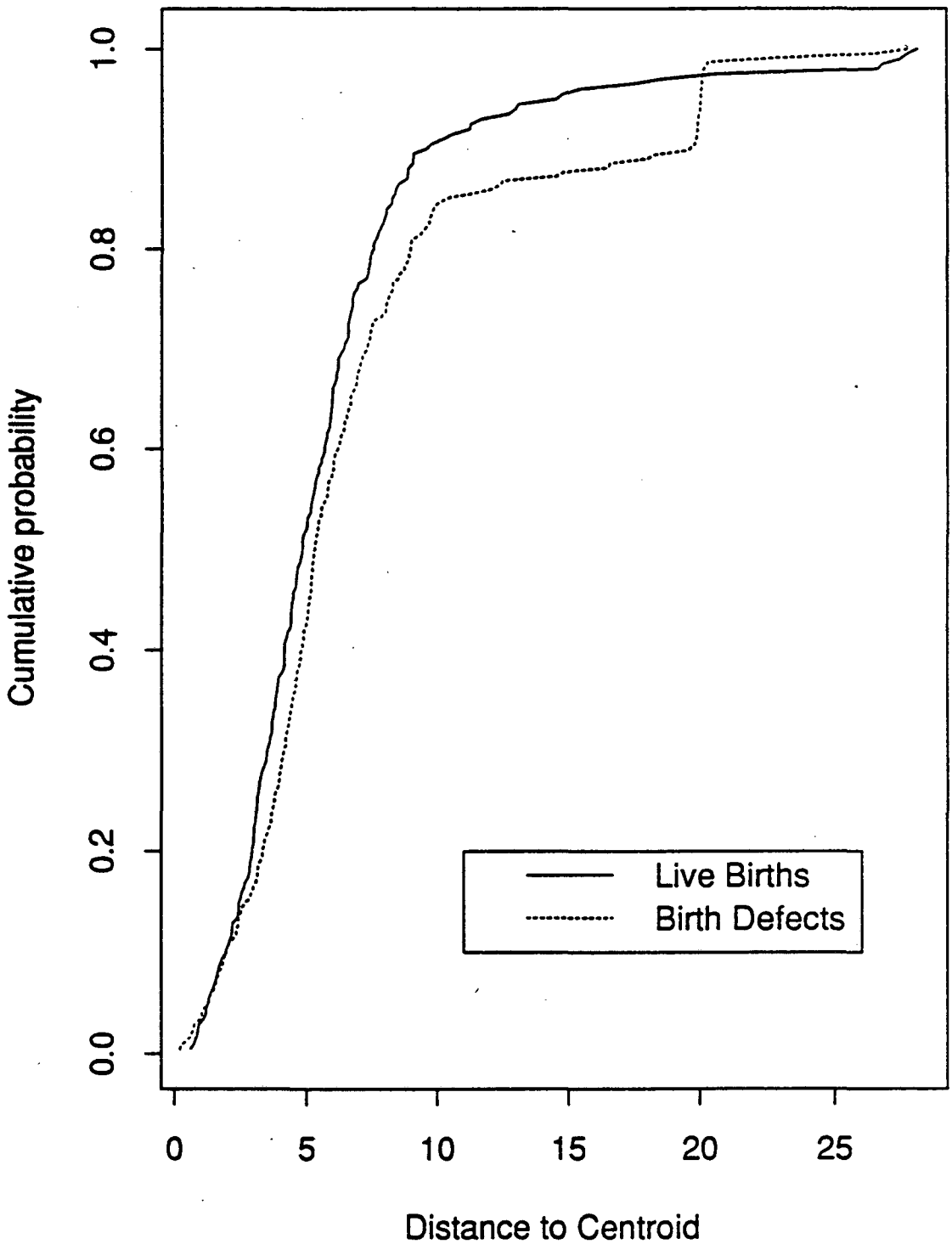


Figure 3.21
Cumulative Probability Distributions
Cluster at D=20 miles



4. Demonstration of Method

4.1. Introduction

In this chapter, the idea of using all birth defects as a surrogate comparison population will be illustrated using a cluster of birth defects that has been previously analyzed and has been shown to exhibit spatial clustering. Background information about this cluster will first be presented, and then a demonstration using three of the seven statistics evaluated in chapter 2 will be presented.

4.2. Background Information on Cluster

Beginning in the Summer of 1986, the California Birth Defects Monitoring Program (CBDMP) began to receive independent reports of a possible cluster of oral cleft defects (Cleft lip with or without Cleft palate) in a neighborhood in San Jose, California. Four children were identified as cases for this cluster, having been born with an oral cleft defect during the years 1980-1986. Three of these four children were born between 1983 and 1986. Since data on the underlying population at risk, and data on all birth defects in the area, were not available for 1980, this re-analysis was restricted to the three children born between 1983 and 1986. In the original investigation, analysis using the baseline data maintained in the CBDMP registry confirmed an estimated 9-fold elevation in the expected prevalence of oral clefts in the neighborhood, based on estimates of the number of births in the neighborhood generated from vital statistics data available for the census tract the neighborhood was in, and the prevalence of oral clefts of 1.9 cases per 1,000 live births, calculated for the San Francisco Bay Area [1]. The excess was statistically significant ($p<0.05$).

All births in the cluster occurred in a 0.2 square mile neighborhood of San Jose. The neighborhood was defined for epidemiologic analysis as the area bounded by the census tract boundary to the west and north (the boundary in this area is a major road), a hilly, mostly uninhabited area to the east, and an uninhabited open space to the south.

Two exposure route allegations, water and air, were made in association with this cluster report. The hypothesized water exposure was to a creek running through the neighborhood which carried discharge water from groundwater clean-up activities occurring at two nearby Environmental Protection Agency "Superfund" sites. While the creek was known to contain some quantity of the organic contaminants 1,1,1-trichloroethane (TCA) and trichlorotrifluoroethane (FREON) this creek was not a known source of drinking water to the neighborhood. The drinking water of all four mothers of the index cases was sampled, and no detectable levels of either TCA or FREON were detected. In addition, the mothers were interviewed as part of the investigation, and all denied having direct contact with the creek.

Allegations of air exposure to sewer odors have been documented in the neighborhood for the previous 20 years. No data were available on the chemical composition of the odors during the period of concern, but air monitoring in the neighborhood for 12 organic constituents showed all values were consistent with ambient levels in the San Francisco Bay Area.

Neither of the two alleged environmental causes was consistent with a "point-source" allegation that could be evaluated with the average distance to a fixed point statistic $\bar{d}$, as contaminant levels in drinking water are not related to distance from the source, and as no point source for the air pollution was identified.

All three of the cases in this cluster exhibited cleft lip, with or without cleft palate (oral clefts), defined as children identified through review of the medical record with British Pediatric Association (BPA) code 749. Other birth defects that occurred in this tract during the period 1983-1986 are shown in table 4.1. Two additional oral cleft cases occured elsewhere in the census tract.

Table 4.1
Birth Defects in Study Tract
1983-1986

Birth Defect	BPA Code	# of cases in Tract
Cleft Lip/Palate	749	6
Down Syndrome	758	2
Prune Belly Syn.	756.72	1
Deformities of Foot	754.7	1
Coronal Synostosis	756.006	1
Deformities of Hip	755.66	1
Congenital Infections	771	1
multiple anomalies	-	1

4.3. Methods

Street addresses were abstracted for all children born with birth defects in this census tract during the years 1983-1986 from the CBDMP, and were manually located and digitized from county assessors maps, as was done for the samples in Chapter 3. Street addresses for all live births occurring during the years 1983-1987 were abstracted from the birth certificates maintained by the county, and were also digitized. It is important to note that the birth dates for cleft lip cases in the cluster, all birth defects in the tract, and live births occurring in the tract do not perfectly overlap. Live births from 1987 were included in the data set because initially it was thought that birth defect records for 1987 would be available before completion of this research from the CBDMP. While birth defects data were not available for

1987, the live birth data was left in, since the spatial distribution of live births did not appreciably change and it was not possible to selectively remove these points without redigitizing all live-birth addresses.

Since no point source exposure was identified that would be appropriate for this investigation, the distance to a fixed point measures d, d^2, and $d_{harmonic}$ were not used to investigate this cluster. The nearest neighbor distance, r, and the average interpoint distance measures, i and i^2 were used. The harmonic average interpoint distance, $i_{harmonic}$ was not calculated as it was found in chapter 2 that no additional information was provided by this statistic, and the decrease in variance compared to i was trivial.

Two sets of analyses were performed. One analysis was limited to the three cluster cases occurring between 1983 and 1986. The second analysis compared the five identified cases of oral clefts in the entire census tract to various comparison groups.

For each of the analyses, two primary comparison groups were used: the set of all birth defects (including the three cluster cases), and the set of all live births (including those born with a birth defect). In addition, for the investigation of the clustered cases of oral clefts, comparisons were also made with the set of all birth defects and the set of live births restricted to those actually in the neighborhood.

When using the randomization or permutation approaches to determining the distribution of the statistic, it is necessary to include the cases from the cluster in the comparison group. The reasoning behind this is that one is trying to determine what the probability is of generating the particular configuration hypothesized to be a cluster by chance. The null hypothesis, that no clustering is present, implies that the configuration of these "clustered" cases occurred by chance. By examining the sorted list of

randomized or permuted spatial statistics, and noting what proportion of them are greater than the observed statistic from the clustered cases, one can establish a "p-value". A summary of the analyses performed appears in table 4.2.

Table 4.2
Analyses Performed

Cases	Comparison Groups
Clustered Oral Clefts 1983-1986 (n=3)	All Birth Defects in Tract 1983-1986 (n=13) All Live Births in Tract 1983-1987 (n=373) All Birth Defects in Neighborhood 1983-1986 (n=5) All Live Births in Neighborhood 1983-1987 (n=143)
All Oral Clefts in Tract 1983-1986 (n=5)	All Birth Defects in Tract 1983-1986 (n=13) All Live Births in Tract 1983-1987 (n=373) All Birth Defects in Tract Except Oral Clefts 1983-1986 (n=8)

As was noted in chapter 2, the interpoint distance and nearest neighbor statistics are both sensitive to the number of cases being evaluated, and thus to compare one population to another using these statistics, it is necessary to insure both statistics are calculated from the same number of points. To accomplish this, permutation and randomization techniques were used.

In the analysis of all three clustered oral cleft cases, all $\binom{13}{3}$ combinations of the 13 birth defects in the tract taken three at a time were generated, and the corresponding statistics were calculated. The distribution of these statistics was then used to establish the level of significance for the observed statistics from the four clustered cases. Similarly all $\binom{5}{3}$ combinations of the five birth defects in the neighborhood were used to generate a distribution based on the neighborhood, and all possible $\binom{8}{3}$ combinations of birth defects in the tract except the oral clefts was used to generate a distribution from the birth defects excluding the defect of interest. A similar

strategy was used for the other case definitions.

Since 373 live births occurred in the tract during the time period available, a randomization approach similar to that used in chapter three was used for these comparisons. For each comparison, 2000 random samples of the number of points in the case group were taken with replacement from the 373 points, and the spatial statistics were calculated. The resulting distributions may be taken to be equivalent to the permutations generated for the birth defects group, above.

4.4. Results—Clustered Oral Cleft Analyses

Maps of the live births in the tract, the three cluster cases, the other birth defects in the tract, and the live births in the neighborhood where the cluster occurred appear as figures 4.1-4.4. The results of the investigation of the three clustered oral cleft cases that occurred between 1983 and 1986 compared to the appropriate comparison groups are shown in table 4.3.

Using tract-level comparison groups, clustering is detected using all three statistics. Using all birth defects or all live births did not appear to make much difference in this analysis, as both show a significant clustering present.

When using neighborhood-level comparison groups, the clustering of the oral cleft cases is not detectable using these statistics. Neither all defects nor all live births produced a p-value of less than 0.05 for any statistic.

Table 4.3
Spatial Analysis of
Three Clustered Oral Clefts (1983-1986)

	r	i	i^2
Observed Values	0.346	0.659	0.622
	p-values		
Comparison Group:			
All Defects in Tract	0.017	0.021	0.028
All Live Births in Tract	0.036	0.055	0.062
All Defects in Neighborhood	0.200	0.200	0.300
All Live Births in Neighborhood	0.074	0.136	0.175

4.5. Results—All Oral Clefts Analyses

A map showing all cases of oral clefts in the census tract appears in figure 4.5. The results of the investigation of all five oral clefts in the tract are presented in table 4.4.

Table 4.4
Spatial Analysis of
Five Oral Clefts in the Tract (1983-1986)

	r	i	i^2
Observed Values	2.209	4.040	22.965
	p-values		
Comparison Group:			
All Defects in Tract	0.565	0.273	0.292
All Live Births in Tract	0.772	0.565	0.577
All Defects Except Clefts in Tract	0.446	0.179	0.268

4.6. Discussion

The reported cluster of oral clefts was easily detected by all three of the statistics used when census-tract level comparison groups were used. A problem with the pemutation approach used with the birth defect comparison groups is that often a small number of permutations are possible, leading to somewhat imprecise p-values. This is particularly evident in the comparison of all defects in the neighborhood to the oral cleft clusters in table 4.3.

Comparison of the clusters to the spatial distribution of live births and all birth defects at the neighborhood level produced non-significant results; oral clefts are no more spatially clustered in the neighborhood than live births or other birth defects are. This must be interpreted in conjunction with the prevalence proportion of the defect in the neighborhood. Thus, while no clustering within the neighborhood was observed, it was known that the prevalence of oral clefts in this neighborhood was nine times what would be expected from baseline data in the San Francisco Bay Area. If one considers an environmental cause for these birth defects, it would not necessarily be expected that these defects would cluster more closely than live births *within* the neighborhood they are in, unless the environmental cause was *so* specific in location that only a few houses were affected. Thus failing to detect clustering within the neighborhood provides additional information on the possible risk factor, since it may be concluded that this risk factor affects the entire neighborhood, and not one specific area.

The detection of clustering within the larger area such as a census tract, or zip code region, may be beneficial in cases where an excess of the disease is not detected based on rates for that geographic area, and the researcher is interested in whether clustering of the defect of interest is occurring *within* the area for which prevalence proporation denominator data are available. It is clear from this research that in such a situation, one can use the spatial distribution of birth defects as a surrogate for the distribution of all live births in determining whether such clustering is occurring.

A sufficient number of birth defects other than the cluster cases must be present in the area to be studied to make use of the spatial distribution of birth defects as a surrogate for all live births. The number necessary will depend on the number of cases in the purported cluster, and is determined

by the size of $\binom{n}{c}$, where n is the number of all birth defects to be used as a comparison, and c is the number of birth defects in the cluster. Table 4.5 shows the minimum detectable p-value for the permutation approach, based on different cluster and comparison group sizes. The data are presented graphically in figure 4.6.

It should be noted that these minimum detectable p-values are based solely on the combination of $\binom{n}{c}$, and it is probably wise to have more than one or two children with defects in addition to those in the cluster for the comparison to be truly valid.

In this chapter, the use of simulation and permutation techniques with spatial statistics has been demonstrated in the reanalysis of a real cluster of birth defects. The use of birth defects as a surrogate population for the true population at risk did not diminish the significance of this cluster, and in general appeared to perform adequately for the assessment of the cluster. It is clear that the size of the cluster and the number of birth defects in the comparison group will determine how sensitive the resulting test can be.

Table 4.5
Minimum Detectable Probabilities
Using Permutation Approach
With Birth Defects as Comparison Group

Number in Cluster	Number of Defects in Comparison Population							
	3	4	5	6	7	8	9	10
3	-	0.250	0.100	0.050	0.029	0.018	0.012	0.008
4	-	-	0.20	0.067	0.029	0.014	0.008	0.005
5	-	-	-	0.166	0.048	0.018	0.008	0.004
6	-	-	-	-	0.143	0.036	0.012	0.005
7	-	-	-	-	-	0.125	0.028	0.008
8	-	-	-	-	-	-	0.111	0.022
9	-	-	-	-	-	-	-	0.100
10	-	-	-	-	-	-	-	-

	11	12	13	14	15	16	17	18
3	0.006	0.005	0.003	0.003	0.002	0.002	0.002	0.001
4	0.003	0.002	0.001	<0.001				
5	0.002	0.001	<0.001					
6	0.002	0.001	<0.001					
7	0.003	0.001	<0.001					
8	0.006	0.002	<0.001					
9	0.018	0.004	0.001	<0.001				
10	0.091	0.015	0.003	<0.001				
11	-	0.083	0.012	0.002	<0.001			
12	-	-	0.077	0.011	0.002	<0.001		
13	-	-	-	0.071	0.009	0.002	<0.001	
14	-	-	-	-	0.067	0.008	0.001	<0.001
15	-	-	-	-	-	0.063	0.007	0.001

4.7. References

1. California Birth Defects Monitoring Program. Unpublished Data. 1990.

Figure 4.1
Live Births in Census Tract (n=373)
Santa Clara County 1983-1987

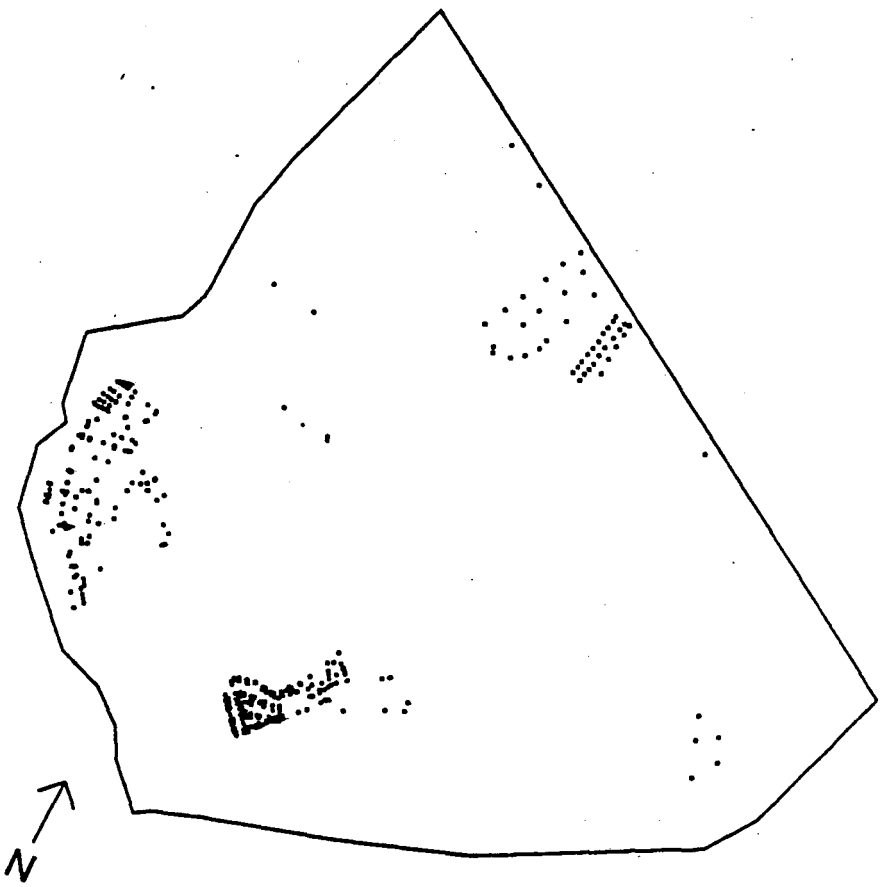


Figure 4.2
Index Cases for Oral Cleft Study (n=3)
Santa Clara County 1980-1986

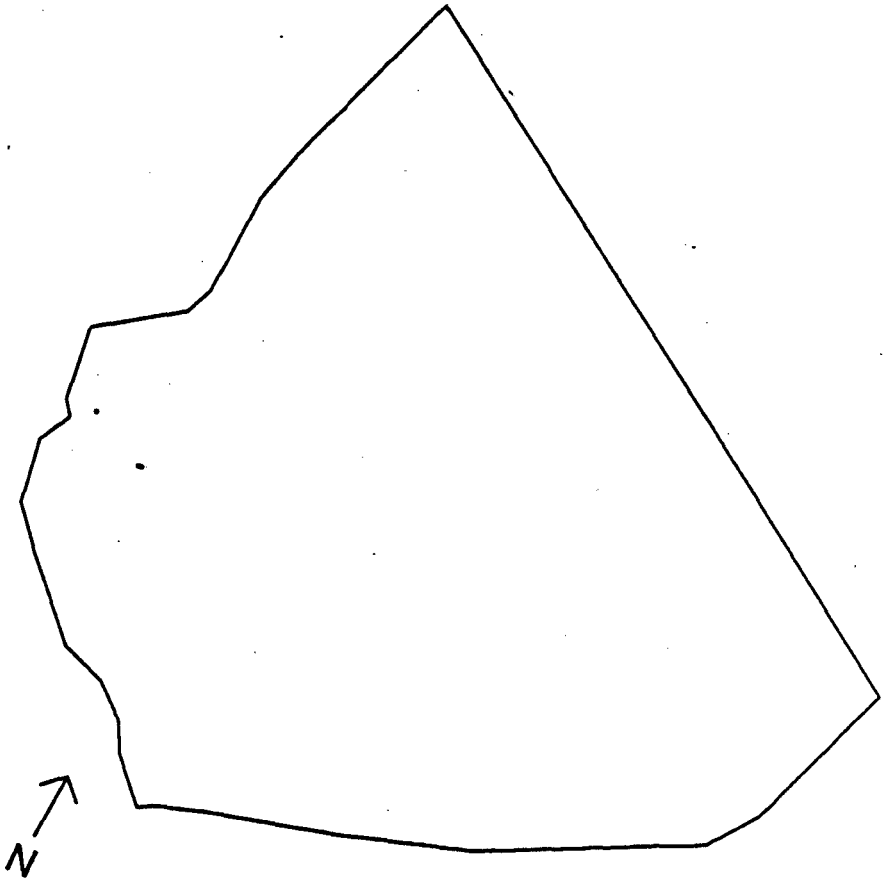


Figure 4.3
All Birth Defects In Tract
for Oral Cleft Study (n=13) 1983-1986

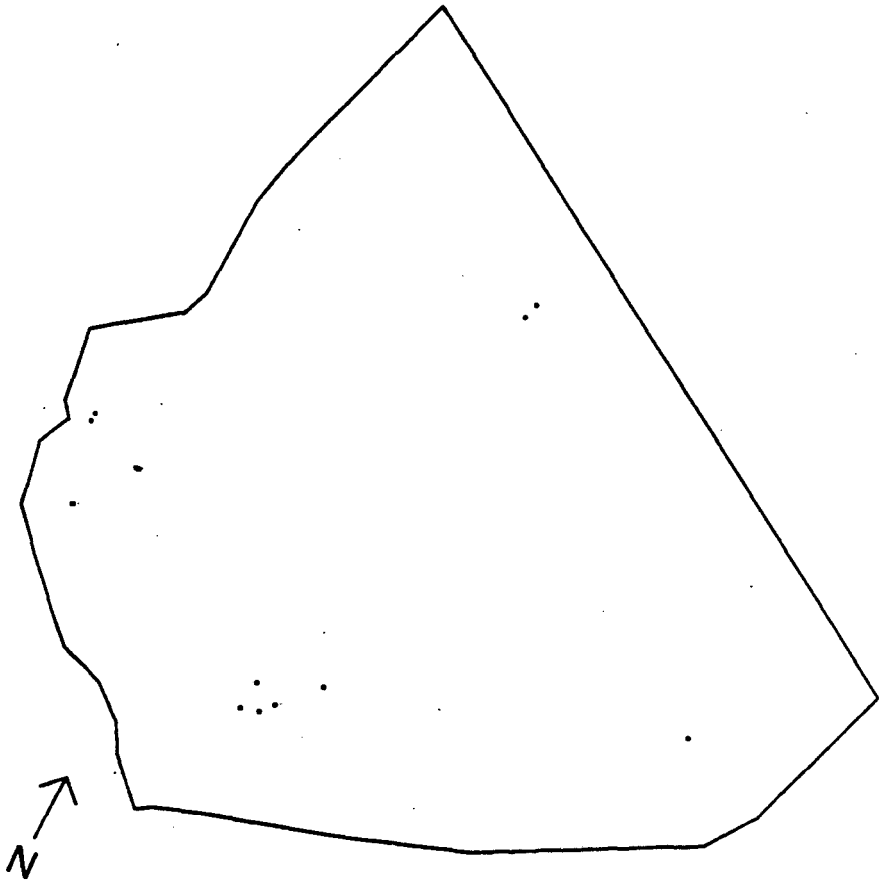


Figure 4.4
Live Births in Neighborhood (n=143)
Santa Clara County 1983-1987

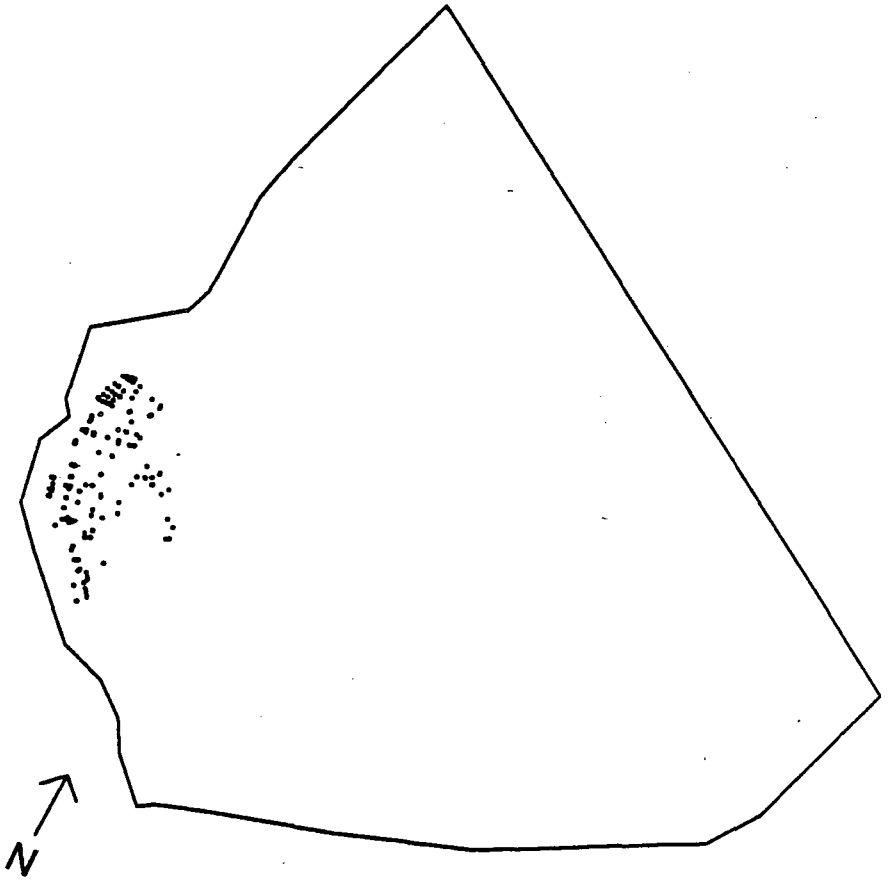


Figure 4.5
All Oral Cleft Births in Tract (n=5)
Santa Clara County 1983-1986

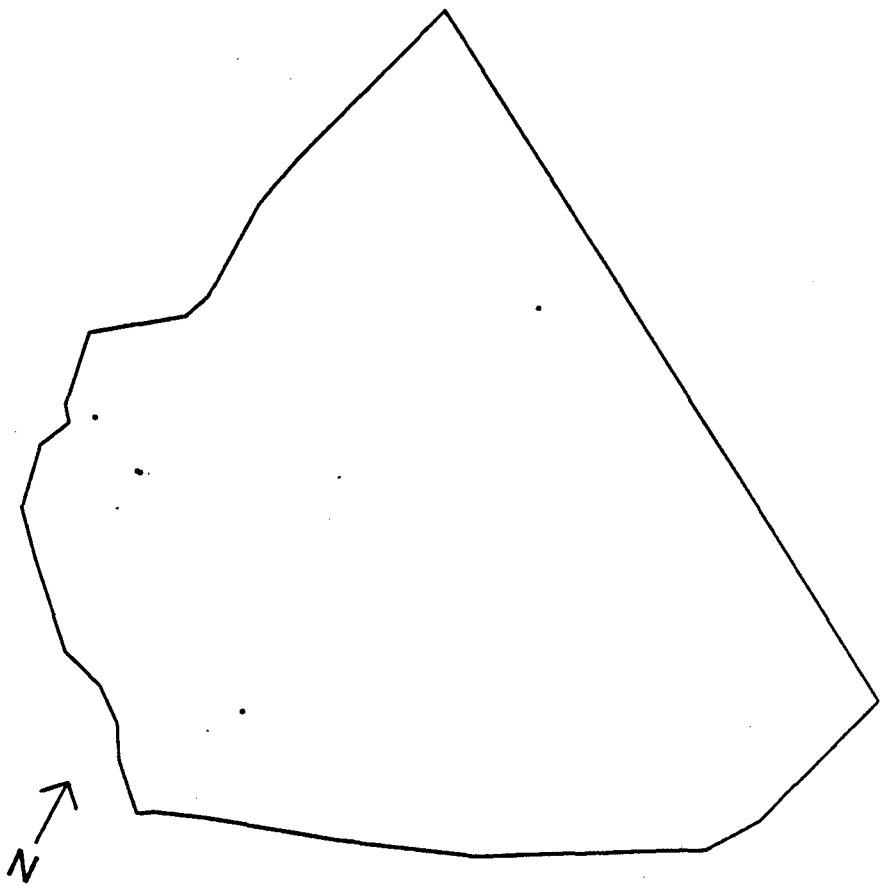
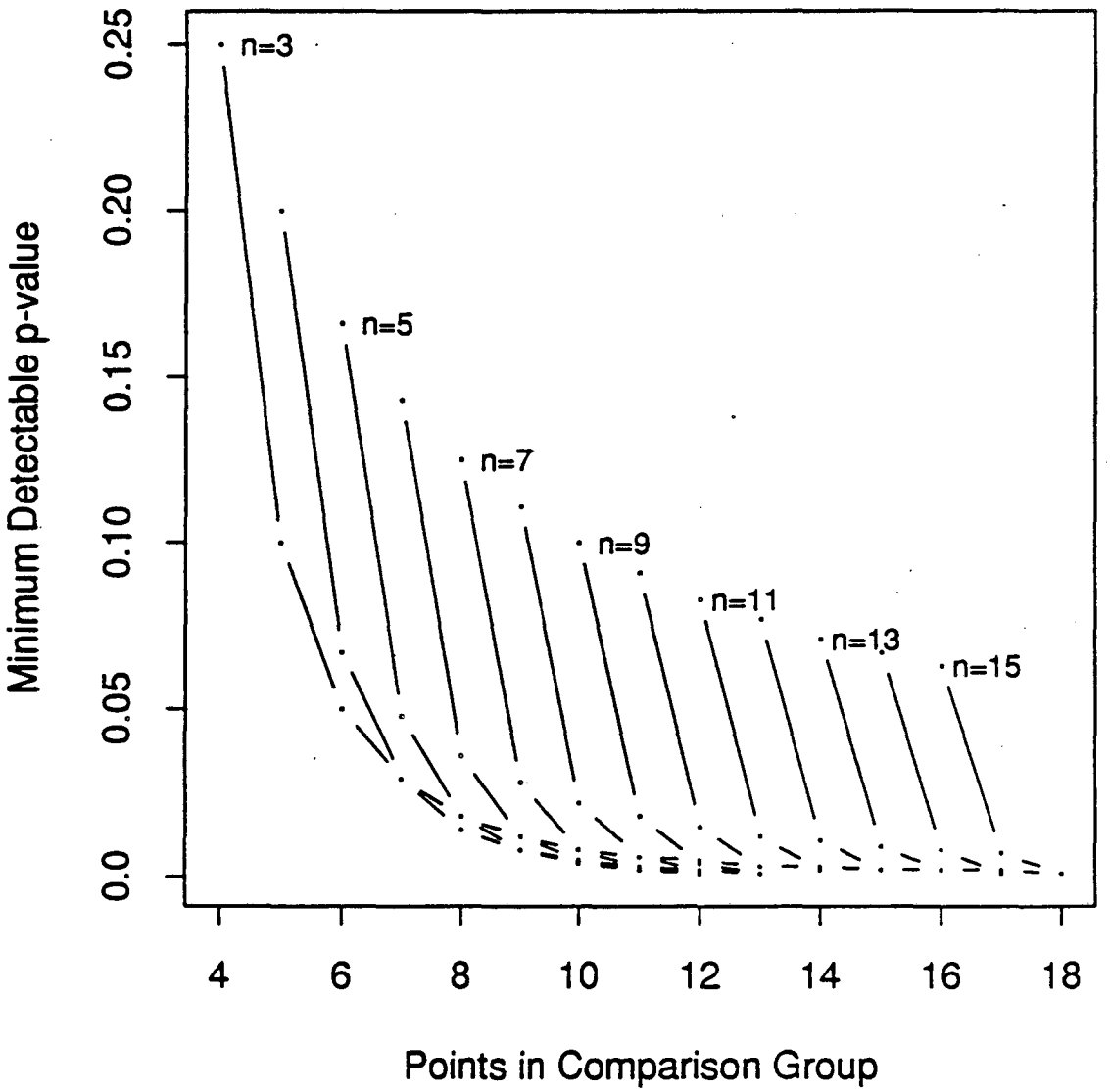


Figure 4.6
Minimum Detectable p-values
using Permutation Approach



5. Epidemiologic Implications

5.1. General Comments

Investigations of clusters serve two fundamental purposes; a social one, in which they assuage community concern over adverse health outcomes from perceived exposures, and a scientific one, in which hypotheses are generated to explain the clustering phenomenon [1, 2]. Both purposes demand a rapid, unbiased response plan, in which the needs of the community are addressed while maintaining standards of scientific rigor. The cost of conducting a full-scale case-control study of a cluster requires a preliminary screening method, such as that described by Grether, et al [3], be used. In this preliminary screening process, basic information that may be readily available to registries or health agencies is used to make an assessment of the significance of the perceived cluster. This assessment will often require the evaluation of the closeness of spatial aggregation of the cluster cases.

5.2. Results of Investigations

All three of the basic spatial statistics investigated were found to be robust in their response to violations of the fundamental assumptions underlying their use. A major concern in the use of the nearest neighbor and average interpoint distance statistics is that the distances that contribute to these means are not completely independent, and thus the assumption of independence is violated. The use of Jackknife or Bootstrap methods for these statistics is therefore not necessary, as they yield equivalent results. The use of harmonic means, as suggested by Mantel [4] resulted in larger variances and it is not clear from this investigation what, if any, benefit is

derived from their use in the spatial analysis of clusters. Indeed, a combination of the harmonic technique and the bootstrap method produces erroneous results, since when the same point is drawn more than once in the same bootstrap sample, the value of the harmonic mean becomes infinite.

The expected values of each of these statistics is derived based on the assumption of uniform distribution of the underlying population at risk. As the underlying population at risk may not be uniformly distributed in space, as demonstrated in chapter 4, the use of these theoretically calculated expected values in the analysis of geopolitical maps should be practiced with caution. A better estimate of the "expected value" of the statistics in a given area may be the values generated from randomization or permutation methods, which result in null distributions from which estimates of p-values for the observed statistic may be determined.

The assessment of whether the population of all birth defects is analogous to the population of all live births performed in chapter 3 attempted to analyse whether the spatial patterns were sufficiently similar to permit their use interchangeably. From this analysis, it was concluded that there is little evidence to suggest that the two populations are fundamentally different, in terms of their spatial distribution. The problem encountered in this analysis is that it is not possible to prove conclusively that two distributions are the same. One can only demonstrate that they are not statistically different according to the metric in use. Since no striking differences between the live births and the birth defects were noted using any of the methods demonstrated, it is probably justifiable to proceed on the assumption that a major bias will not be introduced by using the surrogate population. The use of all birth defects as a surrogate for all live births is further justified, since any clustering of other birth defects in space will introduce a conservative

bias—that is, it will reduce the power to detect a cluster. Hence if a preliminary analysis, using all birth defects as the comparison population, yields evidence of spatial clustering, further efforts to ascertain the spatial distribution of all live births are unwarranted.

In the investigation of a cluster in Santa Clara County, California, presented in Chapter 4, the nearest neighbor distance and the average interpoint distance statistics were applied in the reanalysis of an alleged cluster of oral clefts (cleft lip and cleft palate). Various comparison groups were used to evaluate clustering within the census tract, and within the neighborhood in which the cluster occurred. While clustering was apparant using any of the comparison groups at the census tract level, clustering was not apparent within the neighborhood. This was indicative that, while an excess of oral clefts was occurring in the neighborhood, spatial clustering *within* that neighborhood was not occurring. These results are consistent with either a random cluster, or with an environmental cause that places the entire neighborhood at risk.

One way of putting clusters "in perspective" is to consider how frequently a similar cluster would be expected to occur simply by chance. Since clusters of birth defects occur with small numbers of cases, their distribution can be considered to be Poisson, and an analysis can be performed to determine how frequently a cluster with the number of cases observed would actually occur. For the cluster in chapter 4, 143 births occurred in the neighborhood during the four year period 1983-1986. The baseline rate of oral clefts in the San Francisco Bay Area has been estimated to be 1.9/1000 [5]. Calculating $\hat{\lambda} = np$, we estimate $\hat{\lambda}$ to be 0.2727. Applying the Poisson formula shown in equation 5.1, where k is the number of oral cleft cases in the cluster, we can establish the probabilities for one or more cases of oral

cleft occurring in the neighborhood.

$$P(x=k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{5.1}$$

A registry faced with limited resources to perform in-depth analysis of every cluster report received could make use of this Poisson argument for prioritizing which investigations to perform first. The Poisson method thus may also be a useful adjunct to the spatial methods in the investigation of clusters. From a scientific perspective, a cluster that would be expected only once every 10 years in California may be a greater cause for concern than a cluster such as this one, for which approximately eight such clusters are expected to occur in California each year. Recognizing the dual purposes of cluster investigations, discussed above, some effort must go into the investigation of any cluster. The use of spatial statistics to determine the degree of clustering is one tool that can contribute to such investigations.

Table 5.1
Poisson Distribution of Oral Clefts
in Cluster of 143 births

Number of Cases of Oral Cleft	Probability of Cluster Occurring by Chance
0	0.7621
1	0.2070
2	0.0281
3	0.0025
4	0.0002

5.3. Spatial Methods in Cluster Problems

The use of spatial statistics such as the nearest neighbor distance, average interpoint distance, and distance to a fixed point has the advantage of using continuous data in a continuous sense, rather than dichotomizing it as is done by many other spatial statistics, particularly space-time methods

such as the Knox statistic [6]. A drawback in their use, however, is that the means are not calculated from statistically completely independent values. As was noted in chapter 1, the nearest neighbor distances and the average interpoint distance statistics both lack complete independence. However, the impact of this lack of independence appears to be minimal on their performance, as was seen in chapter 2, and it appears this lack of statistical independence can be disregarded in most studies.

To make use of these continuous spatial statistics in the proper sense, it is necessary to use the exact location of the cases being evaluated. In methods such as the Density Equalized Map Projections of Selvin et al [7], examples have usually made use of the centroid of the census tract as the point location for all cases occurring in that tract. While this technique permits an assessment of clustering over a larger area, such as a city, it cannot serve in the analysis of smaller areas. When street address of the cases is known, digitization techniques can be used to better approximate the distribution of cases and the comparison population. Bias in the digitization is minimal and non-differential with respect to disease status, and provides a rapid means of assigning cartesian coordinates. The use of digitization has the added advantage of maintaining confidentiality of the data, as demonstrated with the maps in this dissertation.

5.4. Implications of Using All Birth Defects as a Surrogate Population

The use of all birth defects as a surrogate for all live births in the spatial investigation of clusters appears justifiable, based on the results of the analyses in chapters 3 and 4. Such use makes the assumption that no "universal teratogen" is modifying the spatial distribution of all defects in space, compared to the distribution of all live births. This assumption is probably

generally valid, provided the defects in the cluster are a small proportion of the total comparison population. It is important that sufficient non-clustered defects be present to prevent the clustered cases from completely "washing out" any observable effect.

Birth defect registries that collect population-based data on defects within a state or county are prime responders to cluster reports. In California, the CBDMP receives, on average, two reports per month of clusters occurring within the state [5]. Such reports are evaluated following an established protocol which includes the spatial analysis of the reported cluster [3]. Such investigation is done to determine whether clustering can be attributed to the spatial distribution of the population, and has often been difficult due to lack of appropriate data describing the spatial location of live births in the area of the cluster. The source of data on live births is usually the county vital statistics office, and often street address is not available. When this is the case, determination of whether clustering is simply due to spatial variation in population density has not been adequately assessed.

Application of the method introduced in this dissertation, using registry data for the comparison group, will permit a more rapid assessment of spatial clustering, and will permit such assessment even when data on the location of live births are not readily available.

It is clear from the demonstrations in Chapters 3 and 4 that, at least in this instance, the spatial distributions of the population of live births and that of all birth defects are similar. Biases that could result in spurious evidence for clustering could occur if the spatial distribution of the comparison group was, in some way, more dispursed in space than the underlying population at risk.

One example of how this could occur can be thought of by referring to figure 4.4. Consider a situation where only seven birth defects occurred in the tract for the study years--the three cluster cases, and one additional defect in each of the four "neighborhoods" of the tract. Under this circumstance, using the four remaining defects as a comparison would be misleading, as the nearest neighbor to each defect would be in another neighborhood, and hence further away than the nearest live birth. In this circumstance, both the nearest neighbor and the average interpoint distance statistics would yield erroneously high values when based on birth defects alone, which, when compared to the cases in the cluster, would make one think strong clustering of the cases was present. To avoid this, one must make sure that either the birth defects are not so sparse in the population that they fail to pick up gross variation in the spatial distribution of live births, or one must verify that no major "clumping" of live births, due to geological or other barriers, is occurring in the study area.

Spatial clustering must be an adjunct to detection of an excess of disease. It is clearly possible to have an excess of disease occurring in some area, such as a census tract, without having "clustering" of the disease. Similarly, it is possible for clustering to be detected in data even though no excess of the disease has occurred. The former situation may be beneficial in helping to eliminate environmental causes that are not plausible for the disease pattern, while the latter situation may indicate that an increased rate of disease is present within a smaller area than was used in the calculation of rates (as was the case in the example in chapter 4). This increase may be due to some environmental cause, or may be a random occurrence. It is clear that when an excess is thought to exist, further examination of the spatial distribution of the cases may yield insight into possible

environmental causes.

5.5. Implications of use of geopolitical boundary

The most commonly used boundaries in epidemiologic investigations are geopolitical, such as census tracts zip code regions and counties. By defining exposure on the basis of these regions, one may misclassify a significant proportion of the cases as exposed, who are actually not exposed to the risk factor of concern. In the example in chapter 4, 2 of the 5 cases of oral cleft that occurred in the tract (40%), would be misclassified if one considered the entire tract "exposed." Thus, some care should go into the definition of the boundary of the cluster. The use of digitization allows greater freedom to select an appropriate boundary. While the use of street address requires additional work to ascertain the location of cases, the benefit is a more useful method of analysis, since with a more precise boundary, the possible environmental and other causes of the cluster can be narrowed down more effectively.

5.6. Epidemiologic utility of this method

Based on the results of these investigations, and based on prior protocols (e.g. [3]) and editorials (e.g. [1]), the following protocol for the investigation of neighborhood or community clusters reported to registries is proposed.

1. **Initiation of investigation through report of index case(s) to registry.** When a cluster is reported to a registry, the investigator first needs to gather information on cases identified by the reporting party to be part of the cluster. The goal of this data collection is to characterize as completely as possible the population at risk, according to the reporting

party, and to insure that reported cases are, in fact, truly cases of birth defects. Information should also be gathered on environmental or other factors thought to be possible causes of the cluster, if any.

2. **Background Information.** Information on environmental and other factors that have been previously suggested as risk factors for the outcome under consideration should be reviewed early in the investigation. It is possible that genetic or other factors may explain, in part or in total, the prevalence of the outcome, and environmental factors may not play a significant role. Examples of such outcomes include many chromosomal defects and fetal alcohol syndrome, for which the risk factors are genetic and behavioral, respectively.

3. **Case Definition.** Based on the results of the initial report, a case definition should be established for the cluster. This case definition should include the precise condition(s) for which clustering is suggested, the time period during which the cluster occurred, the spatial area in which the cluster occurred, and any other data that will help characterize who is a case. To avoid bias, the case definition should be based on standardized methods. The condition should be defined based on commonly used classification such as the BPA. Multiple types of birth defects should be grouped in the cluster only if there is a biological reason for doing so. Otherwise, different defects should be treated as separate in the analyses that follow. The time period and area in which the cluster occurred should be defined, if possible, based on the exposure that is hypothesized to cause the outbreak. If no exposure is hypothesized, efforts should be made to define possible risk factors for the disease, and exposure to these risk factors should form the basis for defining the area and time period at risk. To avoid

spurious results, it is critical for the region of concern and the time period to be defined based on some hypothesis, rather than based on the location and date of birth of the cases. The index case, and other cases initially reported to the registry, should be reviewed carefully to insure they satisfy the case definition. Those who do not completely satisfy the case definition should either be excluded, or treated separately.

4. **Define Boundaries of Neighborhood.** The boundaries of the neighborhood to be considered must be defined in conjunction with the case definition. Boundaries should, if possible, reflect natural breaks in the spatial distribution of live births, such as geophysical features (hills, bodies of water, etc.) or otherwise unpopulated or sparsely populated areas (industrial parks, vacant lots, etc.). If such natural breaks are not available, the investigator may have to resort to using major roads or other, less distinct boundaries. The purpose of careful boundary definition is to insure that those within the boundary are exposed to the environmental factor of concern, while those outside the boundary are not.

5. **Perform Additional Casefinding.** Once the boundaries of the area of concern and the case definition have been established, additional casefinding, using registry data if available, should be undertaken.

6. **Determine Population at Risk.** In addition to finding additional cases of the outcome of concern, a measure of the population at risk must be established. If vital statistics data are available, the investigator may be able to use live births as the comparison population, thereby calculating a direct estimate of the relative risk. Often vital statistics data are not available for the neighborhood in which the alleged cluster is occurring.

When this is the case, rates can be calculated for a larger geopolitical area, such as a census tract, zip code region, or county, and data on all birth defects in the region for the time period of concern can be used in the spatial analysis below to determine whether clustering within this larger area is occurring.

7. **Determine Prevalence of Outcome.** Using the estimate of the population at risk, the prevalence of the defect of concern should be calculated. If this prevalence is in excess of the baseline prevalence from the registry for the same region, further investigation may be warrented. Grether et al [3] recommend a p-value of less than 0.01 before considerable resources are expended on further investigation of the cluster. This seems to be an arbitrary value, and in some circumstances, it may be deemed beneficial to continue the cluster investigation with a larger p-value, particularly if the exposure of concern is well characterized and is biologically plausible as a risk factor.

8. **Spatial Clustering.** Determination of whether clustering is occurring within the neighborhood may provide additional insight as to the relation between environmental factor and defect, or may clarify a relative risk that is not statistically significant in a larger geopolitical area. If a point source of the environmental cause such as a dumpsite or a smokestack is considered, the "distance to a fixed point" statistics are appropriate measures of the clustering. If the exposure source is not a single location, but occurs throughout the neighborhood, the distance to a fixed point is not appropriate, but the average interpoint distance, or alternately the nearest neighbor statistic, will be. It is useful to note that the clustering of defects in the neighborhood, when evaluated in a larger area such as a census tract, provides an assessment of whether an

excess of that particular defect is occurring in that neighborhood. Thus if the population of all birth defects is being used as a surrogate for the live birth population, the spatial analysis of the cluster defects compared to all defects in the tract may provide reassurances on whether a true excess of the defect of concern is present.

9. **Further assessment of the cluster.** If the data analyses above result in a bona fide cluster of defects, with a significant excess of the disease of concern and clustering in a specific area, a true case-control study could be initiated. Biological plausibility, strength of association in the above analyses, and ability to conduct an unbiased study will all contribute to the decision of whether such a case-control study can be performed.

5.7. Summary

The three major goals of this dissertation were to compare three spatial statistics to determine which were most suited to the investigation of spatial clusters, to evaluate whether all birth defects could be used as a surrogate for all live births in the investigation of these clusters, and to demonstrate the spatial evaluation of neighborhood clusters using spatial statistics and all birth defects as a surrogate comparison population.

Comparison of the statistics to their theoretically calculated values in a unit square demonstrated that when clustering was not present, the statistics yielded reasonable approximations of their expected values. It was found that nonparametric estimation using the Bootstrap and Jackknife procedures was not necessary, as the statistics were robust when the assumptions of normality and independence were violated. Calculation of harmonic means, as suggested by Mantel [4] did not yield appreciably better measures of

clustering. It was concluded that comparing calculated statistics for a set of events for which clustering was suspected to a non-clustered set would detect whether clustering was present, provided the two samples were the same size.

Evaluation of whether all birth defects could be used as a surrogate for all live births in the spatial analysis of a cluster showed that the population of all birth defects appears to be randomly distributed among all live births in space for the example presented, and thus such a substitution is probably feasible. Sensitivity tests of the statistics used for this analysis, however, indicated that power to detect clustering was dependent on the distance of the cluster from the centroid of the live birth population, and that small variations would not be detected.

The reanalysis of a known cluster of oral clefts in Santa Clara County demonstrated that using all birth defects as a comparison group, and applying a randomization procedure, was successful at emulating the results obtained by using all live births. Thus, the use of all birth defects as a surrogate for live births may be a viable alternative when data on the population at risk is not available.

In conclusion, this dissertation has presented a possible alternative method for investigating whether clustering of birth defects may be occurring in space. The use of this method will improve epidemiologic investigations of clusters by more accurately defining the spatial area in which the cluster is occurring, thereby reducing misclassification, and will permit the investigation of clusters for which no data are available on the population at risk, provided a population-based registry is collecting suitable data on birth defects in the affected area.

5.8. References

1. Rothman, K. J. Clustering of disease [editorial]. *Am J Public Health.* 77: 13-15; 1987.
2. Schulte, P. A.; Ehrenberg, R. L; Singal, M. Investigation of occupational cancer clusters: theory and practice. *Am J Public Health.* 77: 52-56; 1987.
3. Grether, J. K.; Harris, J. A.; Hexter, A. C.; Jackson, R. J. Investigating clusters of birth defects: guidelines for a systematic approach. Berkeley CA: California Department of Health Services, California Birth Defects Monitoring Program; 1988.
4. Mantel, N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 27: 209-220; 1967.
5. California Birth Defects Monitoring Program. Unpublished Data. 1990.
6. Knox, G. Detection of low intensity epidemicity: application to cleft lip and palate. *Br J Prev Soc Med.* 17: 121-127; 1963.
7. Selvin, S.; Merrill, D.; Schulman, J.; Sacks, S.; Bedell, L.; Wong, L. Transformations of maps to investigate clusters of disease. *Soc Sci Med.* 26: 215-221; 1988.

6. Appendicies

6.1. Appendix A - Software for Chapter 2

6.1.1. Monte.c

Important Variables

Variable	Contains
*fp	pointer to output file
*ofile	pointer to name of output file
BOOTLIM	number of bootstrap replications (max=1000)
COUNT	number of simulated points (max=50)
REPLIM	number of monte carlo simulations to perform
ai2evar	variance of expected average squared interpoint distance
ai2exp	expected value of average squared interpoint distance statistic
aievar	variance of expected average interpoint distance
aiexp	expected value of average interpoint distance statistic
aihevar	variance of expected average harmonic interpoint distance
aihexp	expected value of average harmonic interpoint distance statistic
aipd	observed average interpoint distance statistic
aipd2	observed average squared interpoint distance statistic
aipdhm	observed average harmonic interpoint distance statistic
ba2avg	bootstrap average squared interpoint distance
ba2bias	bootstrap bias of ba2avg
ba2var	bootstrap variance of ba2avg
bahbias	bootstrap bias of baihavg
bahvar	bootstrap variance of baihavg
baiavg	bootstrap average interpoint distance
baibias	bootstrap bias of baiavg
baihavg	bootstrap average harmonic interpoint distance
baivar	bootstrap variance of baiavg
baix[]	x-coordinates of bootstrap sample points
baiy[]	y-coordinates of bootstrap sample points
bd2avg	bootstrap average squared distance to a fixed point
bd2bias	bootstrap bias of bd2avg
bd2var	bootstrap variance of bd2avg
bdavg	bootstrap average distance to a fixed point
bdbias	bootstrap bias of bdavg
bdhbias	bootstrap bias of bdhdavg
bdhdavg	bootstrap average harmonic distance to a fixed point
bdhvar	bootstrap variance of bdhdavg
bdvar	bootstrap variance of bdavg
bravg	bootstrap average nearest neighbor distance
brbias	bootstrap bias of average interpoint distance
brvar	bootstrap variance of average nearest neighbor distance
db2exp	expected value of average squared distance to a fixed point statistic

db2var	variance of expected average squared distance to a fixed point
dbaravg	observed average distance to a fixed point statistic
dbarvar	variance of observed average distance to a fixed point
dbexp	expected value of average distance to a fixed point
dbhexp	expected value of average harmonic distance to a fixed point
dbhmavg	observed average harmonic distance to a fixed point statistic
dbhmvar	variance of observed average harmonic distance to a fixed point
dbhvar	variance of expected average harmonic distance to a fixed point
dbsqravg	observed average squared distance to a fixed point statistic
dbsqrvar	variance of observed average squared distance to a fixed point
dbvar	variance of expected average distance to a fixed point
fixx	x-coordinate of fixed point
fixy	y-coordinate of fixed point
jai2avg	jackknife average squared interpoint distance
jai2bias	bias of jackknife average squared interpoint distance
jai2var	variance of jackknife average squared interpoint distance
jaiavg	jackknife average interpoint distance
jaibias	bias of jackknife average interpoint distance
jaihavg	jackknife average harmonic interpoint distance
jaihbias	bias of jackknife average harmonic interpoint distance
jaihvar	variance of jackknife average harmonic interpoint distance
jaivar	variance of jackknife average interpoint distance
jd	jackknife average distance to a fixed point
jd2	jackknife average squared distance to a fixed point
jd2bias	bias of jackknife average squared distance to a fixed point
jd2var	variance of jackknife average squared distance to a fixed point
jdbias	bias of jackknife average distance to a fixed point
jdh	jackknife average harmonic distance to a fixed point
jdhbias	bias of jackknife average harmonic distance to a fixed point
jdhvar	variance of jackknife average harmonic distance to a fixed point
jdvar	variance of jackknife average distance to a fixed point
jr	jackknife average nearest neighbor distance
jrbias	bias of jackknife average nearest neighbor distance
jrvar	variance of jackknife average nearest neighbor distance

ravg	observed average nearest neighbor distance statistic
rexp	expected value of average nearest neighbor statistic
rpoint	index of randomly selected bootstrap point
rvar	variance of observed average nearest neighbor distance (in function expected, used to mean variance of expected average nearest neighbor distance)
varpd	variance of observed average interpoint distance
varpd2	variance of observed average squared interpoint distance
varpdhm	variance of observed average harmonic interpoint distance
x[]	x-coordinates of simulated points
y[]	y-coordinates of simulated points
z	correction factor for random number generator

```
#include "stdio.h"
#include "math.h"

    /* PROGRAM: monte   WRITTEN: May, 1989
       PURPOSE: This program is the base program for the
       simulation of seven spatial statistics.  This version
       of the program generates simulated points on a unit
       square, then calculates the nearest neighbor distance,
       the average interpoint distance and variations, and the
       the distance to a fixed point and variations.  The
       statistics are calculated three ways, directly, and by
       means of bootstrap and jackknife methods.  The program
       also calculates the expected values of these statistics
       using the formulae of Schulman, 1986.  The output
       of this program is a dataset that can be read into the
       carlo program either by means of a pipe in unix, or by
       creating a (very large) intermediate file by redirection
       of standard output.

       NOTE ON ZIP CODE REGIONS: Modifications to this program were
       done to permit analysing data in a zip code region.  These
       modifications are not included in this dissertation, but
       are easily accomplished by modifying the code that simulates
       the points to read standard input instead.

       NOTE ON CLUSTERS: Modifications to this program were done
       to simulate clusters of cases in the center of the unit
       square.  These modifications are also not included in this
       dissertation, but involve adding a restriction on where some
       proportion of the cases are allowed to be, and continuing to
       regenerate random numbers until this restriction is met.

       FUNCTIONS: mainline, expect, strap, jack, choose, xhypot
    */

/* define globally recognized variables */
int COUNT=4, BOOTLIM=500, REPLIM=1000, reps, z;
float rexp, db2exp, dbexp, dbhexp, aiexp, ai2exp, aihexp,
     ravg, aipd, aipd2, aipdhm, dbaravg, dbsqravg, dbhmavg;
FILE *fp;

/* start main program which runs once only */
main()
{
    char *ofile;
    void mainline(), expect();

    /* define name of output file */
    ofile = "test.out";
    /* establish correction factor for random number generator */
    z=(int)pow(2.,31.)-1.;

    /* open output file */
    if((fp=fopen(ofile,"w"))==NULL){
```

```
        printf("Cannot open output file\n");
        exit(1);
    }

    /* call function expected with centroid 0.5,0.5 */
    expect(0.5,0.5);

    /* seed random number generator */
    srand(getpid());

    /* call function mainline once for each simulation */
    for(reps=0;reps<REPLIM;reps++){
        mainline();
    }

    /* close output file */
    fclose(fp);
}

/*----------------END MAIN PROGRAM----------------*/
/*-------------BEGIN SIMULATION FUNCTION----------------*/

void mainline()
{

    int    ia, ib, ic, id, ie, ig, ih;
    float x[50], y[50],r[50], fixx,fixy,hypo,rvar,
          varpd, varpd2, varpdhm, xhypot(), dx, dy,
          dbar[50], dbarvar, dbsqrvar, dbhmvar;
    void  strap(), jack();

    /* initialize array variables */
    for(ia=0;ia < COUNT; ia++){
        x[ia]=0.0;
        y[ia]=0.0;
        dbar[ia]=0.0;
        r[ia]=0.0;
    }


    /* establish random x and y coordinates in range 0-1 */
    for(ib=0;ib < COUNT;ib++){
     x[ib]=((float)rand()/z);
     y[ib]=((float)rand()/z);
    }

    /* place fixed point at (0.5,0.5) */
    fixx=0.5;
    fixy=0.5;

    /* initialize statistic variables */
    ravg=0.0;       aipd=0.0;
    aipd2=0.0;      aipdhm=0.0;
    dbaravg=0.0;    dbsqravg=0.0;
```

```
dbhmavg=0.0;    rvar=0.0;
dbarvar=0.0;    dbsqrvar=0.0;
dbhmvar=0.0;    varpd=0.0;
varpd2=0.0;     varpdhm=0.0;

ie=0;

/* calculate observed statistics */
for(ic=0;ic < COUNT; ic++){
 r[ic]=9.9;
 dbar[ic]=xhypot(fabs(x[ic]-fixx),fabs(y[ic]-fixy));
 dbaravg=dbaravg+dbar[ic];
 dbsqravg=dbsqravg+pow(dbar[ic],2.0);
 dbhmavg=dbhmavg+(1.0/dbar[ic]);

 for(id=0;id < COUNT; id++){
    dx=fabs(x[ic]-x[id]);
    dy=fabs(y[ic]-y[id]);

    if (id != ic){
       hypo=xhypot(dx,dy);
       if (r[ic] > hypo) r[ic]=hypo;
    }

    if(id > ic){
       hypo=xhypot(dx,dy);
       aipd=aipd+hypo;
       aipd2=aipd2+pow(hypo,2.0);
       if (hypo != 0.0) aipdhm=aipdhm+(1.0/hypo);
       ie++;
    }
 }
 ravg=ravg+r[ic];
}

dbaravg=dbaravg/(float)COUNT;
dbsqravg=dbsqravg/(float)COUNT;
dbhmavg=dbhmavg/(float)COUNT;
ravg=ravg/(float)COUNT;
aipd=aipd/(float)ie;
aipd2=aipd2/(float)ie;
aipdhm=(aipdhm/(float)ie);

/* calculate variances for r and dbar */
for(id=0;id < COUNT;id++){
  rvar=rvar+pow((r[id]-ravg),2.0);
  dbarvar=dbarvar+pow((dbar[id]-dbaravg),2.0);
  dbsqrvar=dbsqrvar+pow((pow(dbar[id],2.0)-dbsqravg),2.0);
}

rvar=rvar/(float)(COUNT*(COUNT-1));
dbarvar=dbarvar/(float)(COUNT*(COUNT-1));
dbsqrvar=dbsqrvar/(float)(COUNT*(COUNT-1));
dbhmavg=1.0/dbhmavg;
```

```
    dbhmvar=((pow(dbhmavg,4.0)/pow(dbaravg,4.0))*dbarvar);


    /* calculate variances for aipd */
    for(ig=0;ig < COUNT; ig++){
     for(ih=0;ih < COUNT; ih++){
        if (ih > ig){
           dx=fabs(x[ig]-x[ih]);
           dy=fabs(y[ig]-y[ih]);
           hypo=xhypot(dx,dy);
           varpd=varpd+pow(fabs(hypo-aipd),2.0);
           varpd2=varpd2+pow(fabs(pow(hypo,2.0)-aipd2),2.0);
        }
     }
    }

    varpd=varpd/(float) (ie*(ie-1));
    varpd2=varpd2/(float)(ie*(ie-1));
    aipdhm=1.0/aipdhm;
    varpdhm=((pow(aipdhm,4.0)/pow(aipd,4.0))*varpd);

    /* output results of observed data to output file */
    fprintf(fp,"%d %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f",
          reps,ravg, aipd, aipd2, aipdhm, dbaravg, dbsqravg,
           dbhmavg);
    fprintf(fp," %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f\n",
          rvar, varpd, varpd2, varpdhm, dbarvar, dbsqrvar,
           dbhmvar);

    /* call bootstrap function with this simulated set */
    strap(r,dbar,x,y);

    /* call jackknife function with this simulated set */
    jack(r,dbar,x,y);

}

/* ----------------END OF FUNCTION MAINLINE---------------- */

/* ----------------BEGIN EXPECTED FUNCTION---------------- */

void expect(fixx,fixy)
float fixx, fixy;
{
    /* FUNCTION: expect   WRITTEN: May 1989
       PURPOSE: This function calculates the expected values of
     the spatial statistics in a unit square with the fixed
     point as passed to it, according to the equations in
     Schulman, 1986, except nearest neighbor, for which the
     equation can be found in Shaw, 1986

     FUNCTIONS: choose
    */
    float  rvar, db2var, sigsqr, dbvar, choose(), aievar, ai2evar,
```

```
    dbhvar, aihevar;

/* calculated expected nearest neighbor */
rexp=1.0/(2.0*sqrt((float)COUNT));
rvar=0.068/pow((float)COUNT,2.0);

/* calculate expected average interpoint distances */
ai2exp=2.0*2.0*((1.0/3.0)-pow((1.0/2.0),2.0));
aiexp=sqrt(ai2exp);

/* the following is the expected variance of the
   average squared interpoint distance statistic */
ai2evar=(1.0/choose(COUNT,2))*( (2.0*((float)COUNT-1)*(2.0/5.0)) +
       (4.0*((float)COUNT-1)*((1.0/9.0)-(1.0/9.0))) +
      (8.0*((float)COUNT-1)*((1.0/12.0)+(1.0/12.0)-(1.0/8.0))) -
      (8.0*((float)COUNT-1)*((1.0/8.0)+(1.0/12.0)+(1.0/12.0))) -
      (2.0*((float)COUNT-3)*(2.0/9.0)) +
      (8.0*(1.0/4.0)*((2.0*((float)COUNT-2)*(1.0/4.0))+(1.0/4.0))) -
      (4.0*((float)(COUNT*2)-3)*((1.0/4.0)-2*((1.0/12.0)+(1.0/12.0)))));

/* variance of expected average interpoint distance */
aievar=ai2evar/pow((2.0*aiexp),2.0);

/* approximate harmonic expected's as direct expecteds */
aihexp=aiexp;
aihevar=aievar;


/* calculate expected average squared distance to a fixed point */
db2exp= (1.0/3.0) + (1.0/3.0) + pow(fixx,2.0) + pow(fixy,2.0) -
    (2*((fixx*(1.0/2.0))+(fixy*(1.0/2.0))));
db2var= (1.0/(float)COUNT)*((1.0/5.0) + (1.0/5.0) +
    (2*((1.0/3.0)*(1.0/3.0))) -
    (4*((fixx*(1.0/4.0))+(fixy*(1.0/4.0)))) +
    (4*pow(fixx,2.0)*((1.0/3.0)-(1.0/4.0))) +
    (4*pow(fixy,2.0)*((1.0/3.0)-(1.0/4.0))) -
    (pow(((1.0/3.0)+(1.0/3.0)),2.0)) +
    (4*((1.0/3.0)+(1.0/3.0))*((fixx*(1.0/2.0))+(fixy*(1.0/2.0)))) +
    (8*fixx*fixy*(((1.0/2.0)*(1.0/2.0))-((1.0/2.0)*(1.0/2.0)))) -
    (4*((fixy*((1.0/3.0)*(1.0/2.0)))+(fixx*((1.0/2.0)*(1.0/3.0))))));


/* calculate expected average distance to a fixed point */
sigsqr=((float)COUNT)*db2var;
dbexp=(sqrt(db2exp))*(1.0-(sigsqr/(8*db2exp*db2exp))-
   ((15.0/128.0)*((sigsqr*sigsqr)/pow(db2exp,4.0))));
dbvar=(1.0/(float)COUNT)*((sigsqr/(4*db2exp)) +
    ((7.0*sigsqr*sigsqr)/(32.0*pow(db2exp,3.0))) -
    ((15.0*pow(sigsqr,3.0))/(512.0*pow(db2exp,5.0))) -
    ((225.0*pow(sigsqr,4.0))/(16384.0*pow(db2exp,7.0))));

/* approximate harmonic expected's as direct expecteds */
dbhexp=dbexp;
dbhvar=dbvar;
```

```
    /* output results to output file */
    fprintf(fp,"E %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f",
          rexp,aiexp,ai2exp,aihexp,dbexp,db2exp,dbhexp);
    fprintf(fp," %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f\n",
          rvar, aievar, ai2evar, aihevar, dbvar, db2var,dbhvar);

}

/* ----------------END EXPECTED FUNCTION---------------- */

/* ----------------BEGIN BOOTSTRAP FUNCTION---------------- */

void strap(r,dbar,x,y)
  float r[50], dbar[50], x[50],y[50]; {

     /* FUNCTION: strap          WRITTEN: May 1989
        PURPOSE: This function performs a bootstrap analysis on
        the spatial statistics using the simulated values from
        function mainline.
     */

   int rpoint, bk, bb, bc, bd, be, bf, bg, bi, bj;
   float bootr[1000], bootdbar[1000], bootdb2[1000],rpinter,
      boothmd[1000], bravg, bdavg, bd2avg, bhdavg,
      brvar, bdvar, bd2var, brbias, bdbias, bd2bias,
      baix[50], baiy[50],baipd[1000],bai2pd[1000],baihm[1000],
      baiavg, bai2avg, baihavg, baivar,ba2var,bahvar,baibias,ba2bias,
      bhyp,xhypot(), bahbias, bdhbias, bdhvar, intrand;;

   /* initialize array variables */
   for(bd=0;bd<1000;bd++){
      bootr[bd]=0.0;       bootdbar[bd]=0.0;
      bootdb2[bd]=0.0;     boothmd[bd]=0.0;
      baipd[bd]=0.0;       bai2pd[bd]=0.0;
      baihm[bd]=0.0;

      if (bd < 50){
       baix[bd]=0.0;
       baiy[bd]=0.0;
      }

   }

   /* initialize statistic variables */
   bravg=0.0;       bdavg=0.0;
   bd2avg=0.0;          bhdavg=0.0;
   baiavg=0.0;      bai2avg=0.0;
   baihavg=0.0;         brvar=0.0;
   baivar=0.0;      ba2var=0.0;
   bahvar=0.0;          bdvar=0.0;
   bd2var=0.0;          bdhvar=0.0;
   brbias=0.0;      baibias=0.0;
   ba2bias=0.0;         bahbias=0.0;
   bdbias=0.0;      bd2bias=0.0;
```

```
bdhbias=0.0;

/* repeat all of the following BOOTLIM times */
for(bk=0;bk<BOOTLIM;bk++){

   /* establish a bootstrap sample with replacement
    and calculate statistics */
   for(bb=0; bb< COUNT; bb++){
      rpinter=(float)rand()/z;
    rpoint=(int)(rpinter*((float)COUNT));
    bootr[bk]=bootr[bk]+r[rpoint];
    bootdbar[bk]=bootdbar[bk]+dbar[rpoint];
    bootdb2[bk]=bootdb2[bk]+pow(dbar[rpoint],2.0);
    boothmd[bk]=boothmd[bk]+(1.0/dbar[rpoint]);
    baix[bb]=x[rpoint];
    baiy[bb]=y[rpoint];
   }

   /* calculate average interpoint distances for bootstrap sample */
   bj=0;
   for(bf=0;bf<COUNT;bf++){
    for(bg=0;bg<COUNT;bg++){

       if (bg>bf) {
          bhyp=xhypot(fabs(baix[bf]-baix[bg]),fabs(baiy[bf]-baiy[bg]));
          baipd[bk]=baipd[bk]+bhyp;
          bai2pd[bk]=bai2pd[bk]+pow(bhyp,2.0);
          if(bhyp != 0){baihm[bk]=baihm[bk]+(1.0/bhyp);}
          bj++;
       }

    }
   }
   bootr[bk]=bootr[bk]/(float)COUNT;
   bootdbar[bk]=bootdbar[bk]/(float)COUNT;
   bootdb2[bk]=bootdb2[bk]/(float)COUNT;
   boothmd[bk]=boothmd[bk]/(float)COUNT;
   baipd[bk]=baipd[bk]/(float)bj;
   bai2pd[bk]=bai2pd[bk]/(float)bj;
   baihm[bk]=baihm[bk]/(float)bj;

   bravg=bravg+bootr[bk];
   bdavg=bdavg+bootdbar[bk];
   bd2avg=bd2avg+bootdb2[bk];
   bhdavg=bhdavg+boothmd[bk];
   baiavg=baiavg+baipd[bk];
   bai2avg=bai2avg+bai2pd[bk];
   baihavg=baihavg+baihm[bk];

}

/* end of bootstrap replications, calculate means */
bravg=bravg/BOOTLIM;
bdavg=bdavg/BOOTLIM;
```

```
    bd2avg=bd2avg/BOOTLIM;
    bhdavg=bhdavg/BOOTLIM;
    baiavg=baiavg/BOOTLIM;
    bai2avg=bai2avg/BOOTLIM;
    baihavg=baihavg/BOOTLIM;


    /* calculate bootstrap variance and bias */
    for(be=0;be<BOOTLIM;be++){
       brvar=brvar+pow((bootr[be]-bravg),2.0);
       baivar=baivar+pow((baipd[be]-baiavg),2.0);
       ba2var=ba2var+pow((bai2pd[be]-bai2avg),2.0);
       bdvar=bdvar+pow((bootdbar[be]-bdavg),2.0);
       bd2var=bd2var+pow((bootdb2[be]-bd2avg),2.0);
       brbias=brbias+(bootr[be]-ravg);
       baibias=baibias+(baipd[be]-aipd);
       ba2bias=ba2bias+(bai2pd[be]-aipd2);
       bdbias=bdbias+(bootdbar[be]-dbaravg);
       bd2bias=bd2bias+(bootdb2[be]-dbsqravg);
       bahbias=bahbias+(baihm[be]-(1.0/aipdhm));
       bdhbias=bdhbias+(boothmd[be]-(1.0/dbhmavg));
    }
    brvar=brvar/(float)(BOOTLIM-1);
    bdvar=bdvar/(float)(BOOTLIM-1);
    bd2var=bd2var/(float)(BOOTLIM-1);
    baivar=baivar/(float)(BOOTLIM-1);
    ba2var=ba2var/(float)(BOOTLIM-1);
    bahvar=(pow((1.0/baihavg),4.0)/pow(baiavg,4.0))*baivar;
    bdhvar=(pow((1.0/bhdavg),4.0)/pow(bdavg,4.0))*bdvar;
    brbias=brbias/(float)BOOTLIM;
    bdbias=bdbias/(float)BOOTLIM;
    bd2bias=bd2bias/(float)BOOTLIM;
    baibias=baibias/(float)BOOTLIM;
    ba2bias=ba2bias/(float)BOOTLIM;
    bdhbias=bdhbias/(float)BOOTLIM;
    bahbias=bahbias/(float)BOOTLIM;

    /* output results to output file */
    fprintf(fp,"%d %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f",
          reps,bravg,baiavg,bai2avg,(1.0/baihavg),bdavg,bd2avg,(1.0/bhdavg));
    fprintf(fp," %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f",
          brvar,baivar, ba2var, bahvar, bdvar, bd2var,bdhvar);
    fprintf(fp," %-7.4f %-7.4f %-7.4f %-7.4f %-7.4f %-7.4f %-7.4f\n",
          brbias,baibias, ba2bias,bahbias,bdbias,bd2bias,bdhbias);

}

/* ----------------END BOOTSTRAP FUNCTION---------------- */
/* ----------------BEGIN JACKNIFE FUNCTION---------------- */

void jack(r,dbar,x,y)
float r[50], dbar[50], x[50], y[50];
{
     /* FUNCTION: jack          WRITTEN: May 1989
```

```
      PURPOSE: This function performs a jackknife analysis of
      the spatial statistics using the simulated data from
      mainline.
   */

int ja, jb, jc, je, jf, jg, jk, jl, jm, jn, jp;
float jravg[50],jdavg[50],jd2avg[50],jdhavg[50],jr,jd,jd2,jdh,
    jrvar,jdvar, jd2var, vfix, jaiavg, jai2avg, jaihavg, jai[50],
    jai2[50],jaih[50], jx[50], jy[50],jaivar, jai2var, jaihvar,
    jdhvar, xhypot(), xh, jrbias, jaibias, jai2bias, jdbias, jd2bias,
    jaihbias, jdhbias;

/* initialize statistic variables */
jr=0.0;          jd=0.0;
jd2=0.0;         jdh=0.0;
jaiavg=0.0;      jai2avg=0.0;
jaihavg=0.0;     jrbias=0.0;
jaibias=0.0;     jai2bias=0.0;
jdbias=0.0;      jd2bias=0.0;
jrvar=0.0;       jdvar=0.0;
jd2var=0.0;      jdhvar=0.0;
jaivar=0.0;      jai2var=0.0;
jaihvar=0.0;

/* initialize array statistics */
for(jc=0;jc<50; jc++){
   jravg[jc]=0.0;     jdavg[jc]=0.0;
   jd2avg[jc]=0.0;    jdhavg[jc]=0.0;
   jx[jc]=0.0;        jy[jc]=0.0;
   jai[jc]=0.0;       jai2[jc]=0.0;
   jaih[jc]=0.0;
}

/* establish jackknife samples and calculate statistics */
for(ja=0;ja<COUNT; ja++){

   je=0;
   for(jb=0;jb<COUNT;jb++){
    if(ja != jb){
       jravg[ja]=jravg[ja]+r[jb];
       jdavg[ja]=jdavg[ja]+dbar[jb];
       jd2avg[ja]=jd2avg[ja]+pow(dbar[jb],2.0);
       jdhavg[ja]=jdhavg[ja]+(1.0/dbar[jb]);
       jx[je]=x[jb];
       jy[je]=y[jb];
       je++;
    }
   }

   jl=0; jp=0;
   for(jg=0;jg<(COUNT-1);jg++){
    for(jk=0;jk<(COUNT-1);jk++){
       if (jk > jg){
          jai[ja]=jai[ja]+xhypot(fabs(jx[jg]-jx[jk]),
```

```
                                fabs(jy[jg]-jy[jk]));
                jai2[ja]=jai2[ja]+pow(xhypot(fabs(jx[jg]-jx[jk]),
                                fabs(jy[jg]-jy[jk])),2.0);
                if (jai[ja]> 0) { jaih[ja]=jaih[ja]+
                        (1.0/(xhypot(fabs(jx[jg]-jx[jk]),
                                fabs(jy[jg]-jy[jk]))));
                                jp++;
                  }
                jl++;
            }
          }
        }

        jravg[ja]=jravg[ja]/(float)(COUNT-1);
        jdavg[ja]=jdavg[ja]/(float)(COUNT-1);
        jd2avg[ja]=jd2avg[ja]/(float)(COUNT-1);
        jdhavg[ja]=1.0/(jdhavg[ja]/(float)(COUNT-1));
        jai[ja]=jai[ja]/(float)jl;
        jai2[ja]=jai2[ja]/(float)jl;
        jaih[ja]=1.0/(jaih[ja]/(float)jp);

        jr=jr+jravg[ja];
        jd=jd+jdavg[ja];
        jd2=jd2+jd2avg[ja];
        jdh=jdh+jdhavg[ja];
        jaiavg=jaiavg+jai[ja];
        jai2avg=jai2avg+jai2[ja];
        jaihavg=jaihavg+jaih[ja];

    }
    jr=jr/(float)COUNT;
    jd=jd/(float)COUNT;
    jd2=jd2/(float)COUNT;
    jdh=jdh/(float)COUNT;
    jaiavg=jaiavg/(float)COUNT;
    jai2avg=jai2avg/(float)COUNT;
    jaihavg=jaihavg/(float)COUNT;

    /* calculate variances */
    for(jf=0;jf<COUNT;jf++){
        jrvar=jrvar+pow((jravg[jf]-jr),2.0);
        jdvar=jdvar+pow((jdavg[jf]-jd),2.0);
        jd2var=jd2var+pow((jd2avg[jf]-jd2),2.0);
    }

    for(jm=0;jm<(COUNT-1);jm++){
        for(jn=0;jn<(COUNT-1);jn++){
          if (jn>jm){
            jaivar=jaivar+pow((xhypot(jx[jm]-jx[jn],
                                y[jm]-jy[jn])-jaiavg), 2.0);
            jai2var=jai2var+pow((pow(xhypot(jx[jm]-jx[jn],
                                jy[jm]-jy[jn]),2.0)-jai2avg),2.0);

          }
```

```
        }
    }

    /* calculate biases */
    jrbias=(jr-ravg);
    jaibias=(jaiavg-aipd);
    jai2bias=(jai2avg-aipd2);
    jdbias=(jd-dbaravg);
    jd2bias=(jd2-dbsqravg);
    jaihbias=(jaihavg-aipdhm);
    jdhbias=(jdh-dbhmavg);
    vfix=(float) (COUNT-1)/COUNT;
    jrvar=jrvar*vfix;
    jdvar=jdvar*vfix;
    jd2var=jd2var*vfix;
    jaivar=jaivar*vfix/(float)(jl*(jl-1));
    jai2var=jai2var*vfix/(float)(jl*(jl-1));

    jaihvar=((pow(jaihavg,4.0)/pow(jaiavg,4.0))*jaivar);
    jdhvar=((pow(jdh,4.0)/pow(jd,4.0))*jdvar);

    /* print output to output file */
    fprintf(fp,"%d %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f",
        reps,jr,jaiavg, jai2avg, jaihavg,jd,jd2,jdh);
    fprintf(fp," %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f %7.5f",
        jrvar,jaivar, jai2var, jaihvar,jdvar,jd2var, jdhvar);
    fprintf(fp," %-7.4f %-7.4f %-7.4f %-7.4f %-7.4f %-7.4f %-7.4f\n",
        jrbias,jaibias, jai2bias, jaihbias, jdbias, jd2bias, jdhbias);

}
/* ----------------END JACKKNIFE FUNCTION---------------------- */
/* ----------------BEGIN UTILITY FUNCTIONS---------------------- */

float xhypot(x,y) float x,y;{

    /* FUNCTION: xhypot          WRITTEN: May 1989
       PURPOSE: Function takes two floating point variables
       corresponding to distance between two points on x and y axis
       and returns a floating point variable of the distance
       between the points.
    */

    float z;
    z=pow(x,2.0)+pow(y,2.0);
    z=sqrt(z);
    return z;
}


float choose(n,m) int n,m;{

    /* FUNCTION: choose  WRITTEN: May 1989
       PURPOSE: Function calculates n choose c. Used by expected
       value function.
```

```
    */
    int fa, fb, fc;
    float nfact, mfact, nmfact;
    float chose;
    nfact=1.0;
    mfact=1.0;
    nmfact=1.0;
    for (fa=1;fa <= n; fa++){
     nfact=nfact*((float)fa);
    }
    for (fb=1;fb<=m; fb++){
     mfact=mfact*((float)fb);
    }
    for (fc=1;fc<=(n-m);fc++){
     nmfact=nmfact*((float)fc);
    }
    chose=nfact/(nmfact*mfact);
    return(chose);
}

/* ----------------END OF UTILITY FUNCTIONS---------------- */
```

6.1.2. Carlo.c

Important Variables

Variable	Contains
*ofile	name of input file
*fp	file pointer
dum1-7	dummy variables for reading in values
aa-ag[]	arrays to hold input variables
xmin	minimum value of statistic
xmax	maximum value of statistic
sum	sum for calculating mean
s	standard deviation
s2	variance
g1	skewness
g2	kurtosis

```
#include <stdio.h>
#include <math.h>
    /* PROGRAM: carlo              WRITTEN: May 1989
       PURPOSE: This program takes the output file from
       program monte and summarizes the results.  It was
       written as a separate program to permit examination
       of the intermediate data, if so desired.

       FUNCTIONS: stats */

/* definition of global variables */
int COUNT=1000;

main()
{
    FILE *fp;
    void  stats();
    char  *ofile, line[150];
    int   rep,i,j;
    float aa[1000],ab[1000],ac[1000],ad[1000],
          ae[1000],af[1000],ag[1000],dum1,dum2,
          dum3,dum4,dum5,dum6,dum7;

       /* open file for reading */
       ofile="test.out";
       if((fp=fopen(ofile,"r"))==NULL){
             printf("cannot open input file");
             exit(1);
       }

       /* initialize array variables */
       for(i=0;i<1000;i++){
             aa[i]=0.0;  ab[i]=0.0;
             ac[i]=0.0;  ad[i]=0.0;
             ae[i]=0.0;  af[i]=0.0;
             ag[i]=0.0;
       }

       /* read in and discard the "expecteds" from
          input file, then read in variables for analysis */
       fgets(line,150,fp); /* expected line */

       for(i=0;i<COUNT;i++){
          fscanf(fp,"%d",&rep);   fscanf(fp," %f",&dum1);
          fscanf(fp," %f",&dum2); fscanf(fp," %f",&dum3);
          fscanf(fp," %f",&dum4); fscanf(fp," %f",&dum5);
          fscanf(fp," %f",&dum6); fscanf(fp," %f",&dum7);

          for(j=0;j<7;j++){
              fscanf(fp," %*f");
          }
          fscanf(fp,"%*d");
          for(j=0;j<21;j++){
              fscanf(fp," %*f");
```

```
    }
    fscanf(fp,"%*d");
    for(j=0;j<21;j++){
        fscanf(fp," %*f");
    }
    aa[i]=dum1;    ab[i]=dum2;
    ac[i]=dum3;    ad[i]=dum4;
    ae[i]=dum5;    af[i]=dum6;
    ag[i]=dum7;

}

/* pass data for analysis by stats function,
   eject page and print titles for each analysis */
printf("observed r\n");
stats(aa,COUNT);

printf("observed i\n");
stats(ab,COUNT);

printf("observed i^2\n");
stats(ac,COUNT);

printf("observed iharm\n");
stats(ad,COUNT);

printf("observed d\n");
stats(ae,COUNT);

printf("observed d^2\n");
stats(af,COUNT);

printf("observed dharm\n");
stats(ag,COUNT);

/* rewind the dataset for the second pass
   to analyse variances*/
rewind(fp);

fgets(line,150,fp); /* expected line */

for(i=0;i<COUNT;i++){
   fscanf(fp,"%d",&rep);
   for(j=0;j<7;j++){
       fscanf(fp," %*f");
   }
   fscanf(fp," %f",&dum1);
   fscanf(fp," %f",&dum2);
   fscanf(fp," %f",&dum3);
   fscanf(fp," %f",&dum4);
   fscanf(fp," %f",&dum5);
   fscanf(fp," %f",&dum6);
   fscanf(fp," %f",&dum7);
```

```
    fscanf(fp,"%*d");
    for(j=0;j<21;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp,"%*d");
    for(j=0;j<21;j++){
        fscanf(fp," %*f");
    }
    aa[i]=dum1; ab[i]=dum2;
    ac[i]=dum3; ad[i]=dum4;
    ae[i]=dum5; af[i]=dum6;
    ag[i]=dum7;


}

/* print titles, pass variances to stat function
   for analysis */
printf("observed r variance\n");
stats(aa,COUNT);
printf("observed i variance\n");
stats(ab,COUNT);
printf("observed i^2 variance\n");
stats(ac,COUNT);
printf("observed iharm variance\n");
stats(ad,COUNT);
printf("observed d variance\n");
stats(ae,COUNT);
printf("observed d^2 variance\n");
stats(af,COUNT);
printf("observed dharm variance\n");
stats(ag,COUNT);

/* rewind dataset for third pass to
   analyse bootstrap values */
rewind(fp);

fgets(line,150,fp); /* expected line */

for(i=0;i<COUNT;i++){
    fscanf(fp,"%*d");
    for(j=0;j<14;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp,"%d",&rep);
    fscanf(fp," %f",&dum1);
    fscanf(fp," %f",&dum2);
    fscanf(fp," %f",&dum3);
    fscanf(fp," %f",&dum4);
    fscanf(fp," %f",&dum5);
    fscanf(fp," %f",&dum6);
    fscanf(fp," %f",&dum7);
    for(j=0;j<14;j++){
        fscanf(fp," %*f");
```

```
    }
    fscanf(fp,"%*d");
    for(j=0;j<21;j++){
        fscanf(fp," %*f");
    }
    aa[i]=dum1; ab[i]=dum2;
    ac[i]=dum3; ad[i]=dum4;
    ae[i]=dum5; af[i]=dum6;
    ag[i]=dum7;

}

printf("bootstrap r\n");
stats(aa,COUNT);
printf("bootstrap i\n");
stats(ab,COUNT);
printf("bootstrap i^2\n");
stats(ac,COUNT);
printf("bootstrap iharm\n");
stats(ad,COUNT);
printf("bootstrap d\n");
stats(ae,COUNT);
printf("bootstrap d^2\n");
stats(af,COUNT);
printf("bootstrap dharm\n");
stats(ag,COUNT);

/* rewind dataset for fourth pass, analysing
   bootstrap variances */
rewind(fp);

fgets(line,150,fp); /* expected line */

for(i=0;i<COUNT;i++){
    fscanf(fp,"%*d");
    for(j=0;j<14;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp,"%d",&rep);
    for(j=0;j<7;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp," %f",&dum1);
    fscanf(fp," %f",&dum2);
    fscanf(fp," %f",&dum3);
    fscanf(fp," %f",&dum4);
    fscanf(fp," %f",&dum5);
    fscanf(fp," %f",&dum6);
    fscanf(fp," %f",&dum7);

    for(j=0;j<7;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp,"%*d");
```

```
    for(j=0;j<21;j++){
        fscanf(fp," %*f");
    }
    aa[i]=dum1; ab[i]=dum2;
    ac[i]=dum3; ad[i]=dum4;
    ae[i]=dum5; af[i]=dum6;
    ag[i]=dum7;

}

printf("bootstrap r variance\n");
stats(aa,COUNT);
printf("bootstrap i variance\n");
stats(ab,COUNT);
printf("bootstrap i^2 variance\n");
stats(ac,COUNT);
printf("bootstrap iharm variance\n");
stats(ad,COUNT);
printf("bootstrap d variance\n");
stats(ae,COUNT);
printf("bootstrap d^2 variance\n");
stats(af,COUNT);
printf("bootstrap dharm variance\n");
stats(ag,COUNT);

/* rewind for fifth pass, analysing
   bootstrap biases */
rewind(fp);

fgets(line,150,fp); /* expected line */

for(i=0;i<COUNT;i++){
    fscanf(fp,"%*d");
    for(j=0;j<14;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp,"%d",&rep);
    for(j=0;j<14;j++){
        fscanf(fp," %*f");
    }
    fscanf(fp," %f",&dum1);
    fscanf(fp," %f",&dum2);
    fscanf(fp," %f",&dum3);
    fscanf(fp," %f",&dum4);
    fscanf(fp," %f",&dum5);
    fscanf(fp," %f",&dum6);
    fscanf(fp," %f",&dum7);

    fscanf(fp,"%*d");
    for(j=0;j<21;j++){
        fscanf(fp," %*f");
    }
    aa[i]=dum1; ab[i]=dum2;
    ac[i]=dum3; ad[i]=dum4;
```

```
    ae[i]=dum5; af[i]=dum6;
    ag[i]=dum7;

  }

  printf("bootstrap r bias\n");
  stats(aa,COUNT);
  printf("bootstrap i bias\n");
  stats(ab,COUNT);
  printf("bootstrap i^2 bias\n");
  stats(ac,COUNT);
  printf("bootstrap iharm bias\n");
  stats(ad,COUNT);
  printf("bootstrap d bias\n");
  stats(ae,COUNT);
  printf("bootstrap d^2 bias\n");
  stats(af,COUNT);
  printf("bootstrap dharm bias\n");
  stats(ag,COUNT);

  /* rewind dataset for fifth pass, analysing
     jackknife values */
  rewind(fp);

  fgets(line,150,fp); /* expected line */

  for(i=0;i<COUNT;i++){
     fscanf(fp,"%*d");
     for(j=0;j<14;j++){
         fscanf(fp," %*f");
     }
     fscanf(fp,"%*d");
     for(j=0;j<21;j++){
         fscanf(fp," %*f");
     }
     fscanf(fp,"%d",&rep);
     fscanf(fp," %f",&dum1);
     fscanf(fp," %f",&dum2);
     fscanf(fp," %f",&dum3);
     fscanf(fp," %f",&dum4);
     fscanf(fp," %f",&dum5);
     fscanf(fp," %f",&dum6);
     fscanf(fp," %f",&dum7);
     for(j=0;j<14;j++){
         fscanf(fp," %*f");
     }
     aa[i]=dum1; ab[i]=dum2;
     ac[i]=dum3; ad[i]=dum4;
     ae[i]=dum5; af[i]=dum6;
     ag[i]=dum7;

  }

  printf("jacknife r\n");
```

```
stats(aa,COUNT);
printf("jacknife i\n");
stats(ab,COUNT);
printf("jacknife i^2\n");
stats(ac,COUNT);
printf("jacknife iharm\n");
stats(ad,COUNT);
printf("jacknife d\n");
stats(ae,COUNT);
printf("jacknife d^2\n");
stats(af,COUNT);
printf("jacknife dharm\n");
stats(ag,COUNT);

/* rewind dataset for sixth pass, analysing
   variances of jacknife statistic */
rewind(fp);

fgets(line,150,fp); /* expected line */

for(i=0;i<COUNT;i++){
    fscanf(fp,"%*d");
    for(j=0;j<14;j++){
       fscanf(fp," %*f");
    }
    fscanf(fp,"%*d");
    for(j=0;j<21;j++){
       fscanf(fp," %*f");
    }
    fscanf(fp,"%d",&rep);
    for(j=0;j<7;j++){
       fscanf(fp," %*f");
    }
    fscanf(fp," %f",&dum1);
    fscanf(fp," %f",&dum2);
    fscanf(fp," %f",&dum3);
    fscanf(fp," %f",&dum4);
    fscanf(fp," %f",&dum5);
    fscanf(fp," %f",&dum6);
    fscanf(fp," %f",&dum7);
    for(j=0;j<7;j++){
       fscanf(fp," %*f");
    }
    aa[i]=dum1; ab[i]=dum2;
    ac[i]=dum3; ad[i]=dum4;
    ae[i]=dum5; af[i]=dum6;
    ag[i]=dum7;

}

printf("jacknife r variance\n");
stats(aa,COUNT);
printf("jacknife i variance\n");
stats(ab,COUNT);
```

```
printf("jacknife i^2 variance\n");
stats(ac,COUNT);
printf("jacknife iharm variance\n");
stats(ad,COUNT);
printf("jacknife d variance\n");
stats(ae,COUNT);
printf("jacknife d^2 variance\n");
stats(af,COUNT);
printf("jacknife dharm variance\n");
stats(ag,COUNT);

/* rewind dataset for seventh and final pass,
   analysing bias in jackknife statistic */
rewind(fp);

fgets(line,150,fp); /* expected line */

for(i=0;i<COUNT;i++){
   fscanf(fp,"%*d");
   for(j=0;j<14;j++){
      fscanf(fp," %*f");
   }
   fscanf(fp,"%*d");
   for(j=0;j<21;j++){
      fscanf(fp," %*f");
   }
   fscanf(fp,"%d",&rep);
   for(j=0;j<14;j++){
      fscanf(fp," %*f");
   }
   fscanf(fp," %f",&dum1);
   fscanf(fp," %f",&dum2);
   fscanf(fp," %f",&dum3);
   fscanf(fp," %f",&dum4);
   fscanf(fp," %f",&dum5);
   fscanf(fp," %f",&dum6);
   fscanf(fp," %f",&dum7);

   aa[i]=dum1; ab[i]=dum2;
   ac[i]=dum3; ad[i]=dum4;
   ae[i]=dum5; af[i]=dum6;
   ag[i]=dum7;

}

printf("jacknife r bias\n");
stats(aa,COUNT);
printf("jacknife i bias\n");
stats(ab,COUNT);
printf("jacknife i^2 bias\n");
stats(ac,COUNT);
printf("jacknife iharm bias\n");
stats(ad,COUNT);
printf("jacknife d bias\n");
```

```
        stats(ae,COUNT);
        printf("jacknife d^2 bias\n");
        stats(af,COUNT);
        printf("jacknife dharm bias\n");
        stats(ag,COUNT);

        /* close file */
        fclose(fp);
}
/*-----------------------END PROGRAM MAIN-----------------------*/
/*----------------------BEGIN FUNCTION STATS----------------------*/
void stats(x,n)
  float x[]; int n;
  {
    /*FUNCTION: stats              WRITTEN: May 1989
     ORIGIN:   originally written by S. Selvin in Fortran
     PURPOSE:  this function calculates various summary statistics
     for an array passed to it of length n and generates a histogram
    */

    float xmin, xmax, itab[50], sum, xbar, s, s2,
          s3, s4, dx, temp, g1, g2, imax, ifact,
          xl, xr;
    int   i, ii, j, k, ntab[50], iscale, npts;

    /* initialize variables for histogram */
    iscale=30; k=10;

    /* establish and print general statistics */
    printf("General Statistics\n");
    for(i=0;i<k;i++){
      ntab[i]=0;
    }

    /* calculate minimum, maximum, and mean */
    xmin=x[1];
    xmax=x[1];
    sum=0.0;
    for(i=0;i<n;i++){
     if(xmin>x[i]){
           xmin=x[i];
     }
     if(xmax<x[i]){
           xmax=x[i];
     }
      sum=sum+x[i];
    }
    xbar=sum/(float)n;

    /* calculate variance, standard deviation,
      skewness and kurtosis */
    s2=0.0;
    s3=0.0;
    s4=0.0;
```

```
dx=(xmax-xmin)/(float)k;
for(i=0;i<n;i++){
   temp=x[i]-xbar;
   s2=s2+pow(temp,2.0);
   s3=s3+pow(temp,3.0);
   s4=s4+pow(temp,4.0);
   for(ii=1;ii<=k;ii++){
      if(x[i]>(ii*dx+xmin)) continue;
      ntab[ii-1]=ntab[ii-1]+1;
      break;
   }
}
if(s2<=0.0){
    printf("  ***** variance=0.0 *****\n");
    exit(0);
}
s2=s2/(n-1);
s=sqrt(s2);
s3=s3/(n-1);
s4=s4/(n-1);
g1=s3/pow(s,3.0);
g2=(s4/pow(s,4.0))-3.0;

printf(" nobs = %d\n",n);
printf("mean = %f\n",xbar);
printf("s2 = %f\n",s2);
printf("s.d. = %f\n",s);
printf("g1 = %f\n",g1);
printf("g2 = %f\n",g2);
printf("min = %f\n",xmin);
printf("max = %f\n",xmax);

/* generate histogram */
imax=0.0;
for(i=0;i<k;i++){
   if(ntab[i]<imax) continue;
   imax=ntab[i];
}
ifact=imax/iscale+1;
for(i=0;i<k;i++){
   itab[i]=ntab[i]/ifact+1;
}
printf(" -------- frequency tabulation -------\n0);
printf("interval frequency histogram\n");
xl=xmin;
for(i=0;i<k;i++){
   xr=xl+dx;
   npts=itab[i];
   printf(" %8.5f to %8.5f        %8d   ",xl,xr,ntab[i]);
   for(j=0;j<npts;j++){
      printf("*");
   }
   printf("\n");
   xl=xr;
```

```
    }
    printf("\n %5f observations approximately per *0,ifact);
}

/*-----------------------END FUNCTION STATS----------------------*/
```

6.2. Appendix B - Software for Chapter 3

6.2.1. Squish.c

Important Variables

Variable	Contains
N	Number of points to be read in
ADD	Number of points to be added to create cluster
RADIUS	Radius of area in which cluster points are added
XMEAN	X-coordinate of center of cluster
YMEAN	Y-coordinate of center of cluster
x[]	x-coordinates of input data
y[]	y-coordinates of input data
newx	x-coordinate of new "clustered" point
newy	y-coordinate of new "clustered" point
theta	polar angle coordinate of new "clustered" point
newr	polar radius coordinate of new "clustered" point
z	correction factor for random number generator

```
# include <math.h>
# include <stdio.h>

    /* PROGRAM: squish    WRITTEN: December, 1989
       PURPOSE: Reads in a series of x,y coordinates of points
       in Santa Clara County, and then generates a cluster by
       adding additional points in a specified area.  Outputting
       all data to standard output.

       FUNCTIONS: rectx, recty */
main()
{

       float x[300], y[300], rectx(),recty(),
             newx, newy, newr, theta, RADIUS, XMEAN,YMEAN;
       int ia, ib, ic, id, dum1, dum2, z, N, ADD;

       /* define parameters for program */
       N=216; ADD=20;
       RADIUS=1.491, XMEAN=94.733, YMEAN=299.8233;

       /* seed random number and establish correction factor */
       srand(getpid());
       z=(int)pow(2.,31.)-1.;

       /* initialize array variables */
       for(ia=0;ia<300;ia++){
             x[ia]=0.0;
             y[ia]=0.0;
       }

       /* read in original x,y coordinates */
       for(ib=0;ib<N;ib++){
             scanf("%d %f %f %d", &dum1, &x[ib], &y[ib], &dum2);
       }

       /* re-output input points with identifiers */
       for(ic=0;ic<N;ic++){
          printf("     %d   %7.3f   %7.3f    2\n",
                ic+202,x[ic],y[ic]);
       }

       /* establish new points in cluster and output
          points are established by calculating their
          polar coordinates and then converting to
          cartesian coordinates */
       for(id=0;id<ADD;id++){
          newr=(float)((rand()/(float)z)*RADIUS);
          theta=(float)((rand()/(float)z)*(2.*3.14159265));
          newx=rectx(newr,theta)+XMEAN;
          newy=recty(newr,theta)+YMEAN;
          printf("     %d   %7.3f   %7.3f    2\n",
                id+ic+202,newx,newy);
       }
```

```
/* ----------------END OF PROGRAM MAIN---------------------- */

/* ----------------BEGIN UTILITY FUNCTIONS---------------------- */
}

float rectx(r,theta)
float r,theta;
{
        /* FUNCTION: rectx   WRITTEN: December 1989
           PURPOSE: This function takes the polar coordinates of
           a point and returns the corresponding cartesian "x"
           coordinate */

        float x;
        double cos();
        x=r*((float)cos(theta));
        return x;
}

float recty(r,theta)
float r,theta;
{
        /* FUNCTION: recty   WRITTEN: December 1989
           PURPOSE: This function takes the polar coordinates of
           a point and returns the corresponding cartesian "y"
           coordinate */

        float y;
        double sin();
        y=r*((float)sin(theta));
        return y;
}

/* ----------------END OF UTILITY FUNCTIONS---------------- */
```

6.2.2. Nneigh4.c

Important Variables

Variable	Contains
C1	number of points in group 1
C2	number of points in group 2
COUNT	total number of points
r1[]	nearest neighbor within group 1
r2[]	nearest neighbor within group 2
r12[]	nearest neighbor in group 2 of point in group 1
r21[]	nearest neighbor in group 1 of point in group 2
x[]	x coordinates input from standard input
y[]	y coordinates input from standard input
group[]	group membership of (x,y) coordinate read in
g	indicator for placing distance in appropriate cell
contin[]	data for contingency table
a	mean distance to point in same group
b	mean distance to point in opposite group

```
#include "stdio.h"
#include "math.h"

main()  /* PROGRAM: nneigh4  WRITTEN: January 10, 1990
           PURPOSE: Program generates nearest neighbor
              data for within and between group stats
              to perform contingency table analysis and
              to generate the coefficient of spatial association
              as described in Sorenson

           SUBROUTINES: None
        */
{
    int ia, ib, COUNT, C1, C2,group[500],j, k, dum1,g, contin[4];
    float x[500],y[500],r1[250], mintst,
        r2[250],r12[250],r21[250],r,a,b;

    COUNT=437; C1=201; C2=236;

    /* initialize array variables */
    a=0.0; b=0.0;
    for(ia=0;ia<500;ia++){
      if(ia<250){
         r1[ia]=0.0;
         r2[ia]=0.0;
         r12[ia]=0.0;
         r21[ia]=0.0;
      }
      x[ia]=0;
      y[ia]=0;
      group[ia]=0;
    }

    /* read in coordinates and group membership */
    for(ib=0;ib<COUNT;ib++){
        scanf("%d %f %f %d",&dum1, &x[ib], &y[ib], &group[ib]);
    }

    /* establish data for contingency table */
    for(j=0;j<COUNT;j++){
      r=99999.9;
      for (k=0;k<COUNT;k++){
            if(k != j){
               mintst=sqrt(((x[k]-x[j])*(x[k]-x[j]))+
                        ((y[k]-y[j])*(y[k]-y[j])));
              if (mintst < r){
                  r=mintst;
                  if (group[j]==1){
                     if (group[k]==1) g=0;
                    if (group[k]==2) g=2;
                  }
                  if (group[j]==2){
                     if (group[k]==2) g=3;
                    if (group[k]==1) g=1;
```

```
                }
            }
        }
    }

    contin[g]++;
}

/* print results of analysis */
printf("Contingency Table Analysis of Nearest Neighbor\n\n");
printf("                    BASE GROUP\n");
printf("              Group 1          Group 2\n");
printf("N.N. Group\n");
printf("   Group 1      %d                %d\n",contin[0],contin[1]);
printf("   Group 2      %d                %d\n",contin[2],contin[3]);
printf("0);

/* calculate distance for cell a of table */
for (j=0;j<C1;j++){
  r1[j]=99999.9;
  for (k=0;k<C1;k++){
        if (k != j){
           mintst=sqrt(((x[k]-x[j])*(x[k]-x[j]))+
                       ((y[k]-y[j])*(y[k]-y[j])));
          if (mintst < r1[j]) r1[j]=mintst;
        }
  }
  a=a+r1[j];
}

/* calculate distance for cell d of table */
for (j=0;j<C2;j++){
  r2[j]=99999.9;
  for (k=0;k<C2;k++){
        if (k != j){
           mintst=sqrt(((x[k+C1]-x[j+C1])*(x[k+C1]-x[j+C1]))+
                       ((y[k+C1]-y[j+C1])*(y[k+C1]-y[j+C1])));
          if (mintst < r2[j]) r2[j]=mintst;
        }
  }
  a=a+r2[j];

}

/* calculate distance for cell b of table */
for (j=0;j<C1;j++){ /* establish r for group 12 */
  r12[j]=99999.9;
  for (k=0;k<C2;k++){
           mintst=sqrt(((x[k+C1]-x[j])*(x[k+C1]-x[j]))+
                       ((y[k+C1]-y[j])*(y[k+C1]-y[j])));
          if (mintst < r12[j]) r12[j]=mintst;
  }
  b=b+r12[j];
}
```

```
        /* calculate distance for cell c of table */
        for (j=0;j<C2;j++){
          r21[j]=99999.9;
          for (k=0;k<C1;k++){
                    mintst=sqrt(((x[k]-x[j+C1])*(x[k]-x[j+C1]))+
                                ((y[k]-y[j+C1])*(y[k]-y[j+C1])));
                   if (mintst < r21[j]) r21[j]=mintst;
          }
          b=b+r21[j];
        }

        /* calculate variables for coeff. of spatial association */
        a=a/COUNT;
        b=b/COUNT;

        printf("COEFFICIENT OF SPATIAL ASSOCIATION ANALYSIS\n\n");
        printf("Transformed Cs = %f\n",((a-b)/(a+b)));

}
```

6.3. Appendix C - Software for Chapter 4

6.3.1. stats.c

Important Variables

Variable	Contains
x[]	x-coordinates of input dataset
y[]	y-coordinates of input dataset
COUNT	number of points in input dataset
fixx	x-coordinate of fixed point
fixy	y-coordinate of fixed point
ravg	average nearest neighbor distance
aipd	average interpoint distance
aipd2	average squared interpoint distance
dbaravg	average distance to fixed point
dbsqravg	average squared distance to a fixed point
dbhmavg	average harmonic distance to a fixed point
rvar	variance of average nearest neighbor distance
varpd	variance of average interpoint distance
varpd2	variance of average squared interpoint distance
dbarvar	variance of distance to fixed point
dbsqrvar	variance of average squared distance to fixed point
dbhmvar	variance of average harmonic distance to fixed point

```
#include <stdio.h>
#include <math.h>

      /* PROGRAM: stats   WRITTEN: February 1990
         PURPOSE: Program generates observed nearest neighbor
         average interpoint distance and distance to a fixed
         point statistics for input data sets.

         FUNCTIONS: xhypot */

main()
{

    int    ia, ib, ic, id, ie, ig, COUNT;
    float x[500], y[500],r[500], fixx, fixy, hypo, rvar,
          varpd, varpd2, varpdhm, xhypot(), dx, dy,
          dbar[500], dbarvar, dbsqrvar, dbhmvar,
          ravg, aipd, aipd2, dbaravg, dbsqravg, dbhmavg;
    char *title;

    title = "SPATIAL ANALYSIS OF ALL LIVE BIRTH DATA";

    /* establish number of input data points */
    COUNT=343;

    /* initialize array variables */
    for(ia=0;ia < COUNT; ia++){
        x[ia]=0.0;
        y[ia]=0.0;
        dbar[ia]=0.0;
        r[ia]=0.0;
    }

    /* read in points for analysis */
    for(ib=0;ib<COUNT;ib++){
        scanf("%f %f", &x[ib], &y[ib]);
    }

    /* establish fixed point as reference for d-bar stats */
    fixx=101.38;  fixy=288.48;

    /* initialize other statistics */
    ravg=0.0;         aipd=0.0;
    aipd2=0.0;        dbaravg=0.0;
    dbhmavg=0.0;      dbsqravg=0.0;

    /* calculate statistics */
    ie=0;
    for(ic=0;ic < COUNT; ic++){
      r[ic]=99999999.9;

      dbar[ic]=xhypot(fabs(x[ic]-fixx),fabs(y[ic]-fixy));
      dbaravg=dbaravg+dbar[ic];
      dbsqravg=dbsqravg+pow(dbar[ic],2.0);
```

```
    dbhmavg=dbhmavg+(1.0/dbar[ic]);

    for(id=0;id < COUNT; id++){
      dx=fabs(x[ic]-x[id]);
      dy=fabs(y[ic]-y[id]);
      if (id != ic){
        hypo=xhypot(dx,dy);
        if (r[ic] > hypo) r[ic]=hypo;
       }
       if(id > ic){
        hypo=xhypot(dx,dy);
        aipd=aipd+hypo;
        aipd2=aipd2+pow(hypo,2.0);
        ie++;
       }
     }
     ravg=ravg+r[ic];
}

dbaravg=dbaravg/(float)COUNT;
dbsqravg=dbsqravg/(float)COUNT;
dbhmavg=dbhmavg/(float)COUNT;
ravg=ravg/(float)COUNT;
aipd=aipd/(float)ie;
aipd2=aipd2/(float)ie;

/* initialize variances */
rvar=0.0;
dbarvar=0.0;
dbsqrvar=0.0;
dbhmvar=0.0;
varpd=0.0;
varpd2=0.0;
varpdhm=0.0;

/* calculate variances for nearest neighbor and
   average distance to a fixed point statistics */
for(id=0;id < COUNT;id++){
    rvar=rvar+pow((r[id]-ravg),2.0);
    dbarvar=dbarvar+pow((dbar[id]-dbaravg),2.0);
    dbsqrvar=dbsqrvar+pow((pow(dbar[id],2.0)-dbsqravg),2.0);
}
rvar=rvar/(float)(COUNT*(COUNT-1));
dbarvar=dbarvar/(float)(COUNT*(COUNT-1));
dbsqrvar=dbsqrvar/(float)(COUNT*(COUNT-1));
dbhmavg=1.0/dbhmavg;
dbhmvar=((pow(dbhmavg,4.0)/pow(dbaravg,4.0))*dbarvar);

/* calculate variances for average interpoint
   distance statistics */
for(ig=0;ig < COUNT; ig++){
  for(ih=0;ih < COUNT; ih++){
     if (ih > ig){
        dx=fabs(x[ig]-x[ih]);
```

```
                dy=fabs(y[ig]-y[ih]);
                hypo=xhypot(dx,dy);
                varpd=varpd+pow(fabs(hypo-aipd),2.0);
                varpd2=varpd2+pow(fabs(pow(hypo,2.0)-aipd2),2.0);
              }
            }
          }
          varpd=varpd/(float) (ie*(ie-1));
          varpd2=varpd2/(float)(ie*(ie-1));

          /* print report of results */
          printf("%s\n0,title);
          printf("Nearest Neighbor Statistics:\n r = %f variance = %f0,
                ravg, rvar);
          printf("Average Distance to Fixed Point (%f,%f) Statistics\n",
                fixx,fixy);
          printf("d = %f  variance = %f\n",dbaravg,dbarvar);
          printf("d^2 = %f  variance = %f\n",dbsqravg,dbsqrvar);
          printf("dharm = %f  variance = %f\n0,dbhmavg,dbhmvar);
          printf("Average Interpoint Distance Statistics\n");
          printf("i = %f  variance = %f\n",aipd,varpd);
          printf("i^2 = %f  variance = %f\n",aipd2,varpd2);

}

/* ----------------END OF FUNCTION MAIN---------------------- */

/* ----------------BEGIN UTILITY FUNCTIONS------------------- */

float xhypot(x,y) float x,y;{

   /* FUNCTION: xhypot          WRITTEN: May 1989
      PURPOSE: Function takes two floating point variables
      corresponding to distance between two points on x and y axis
      and returns a floating point variable of the distance
      between the points.
   */

   float z;
   z=pow(x,2.0)+pow(y,2.0);
   z=sqrt(z);
   return z;
}
/* ----------------END OF UTILITY FUNCTIONS------------- */
```

6.3.2. perm.c

Important Variables

Variable	Contains
x[]	x coordinates of input dataset
y[]	y coordinates of input dataset
COUNT	number of points in input dataset
xsamp[]	x coordinates of sample of 4
ysamp[]	y coordinates of sample of 4
ravg	average nearest neighbor distance
aipd	average interpoint distance
aipd2	average squared interpoint distance

```
#include "stdio.h"
#include "math.h"

    /* PROGRAM: perm   WRITTEN: February 22, 1990
       PURPOSE: Program generates nearest neighbor and average
       interpoint distance statistics for all permutations of 4
       of all birth defects occurring in tract 5031.04.

       FUNCTIONS: xhypot()
    */

main()
{

    int    ia, ib, ic, id, ie,ig,ih,ij,ik, COUNT;
    float x[14], y[14],r[4], hypo,xsamp[4],ysamp[4],
          xhypot(), dx, dy, ravg, aipd, aipd2 ;

    /* initialize array variables */
    for(ia=0;ia < COUNT; ia++){
        x[ia]=0.0;
        y[ia]=0.0;
        if(ia<4){
            xsamp[ia]=0.0;
            ysamp[ia]=0.0;
            r[ia]=0.0;
        }
    }

    /* set the number of points to be read into the program */
    COUNT=14;

    /* read in the x and y coordinates from standard input */
    for(ib=0;ib<COUNT;ib++){
        scanf("%f %f", &x[ib], &y[ib]);
    }

    /* generate all samples of size 4 */
    for(ig=0;ig<14;ig++){
      for(ih=ig+1;ih<14;ih++){
        for(ij=ih+1;ij<14;ij++){
            for(ik=ij+1;ik<14;ik++){

                xsamp[0]=x[ig]; ysamp[0]=y[ig];
                xsamp[1]=x[ih]; ysamp[1]=y[ih];
                xsamp[2]=x[ij]; ysamp[2]=y[ij];
                xsamp[3]=x[ik]; ysamp[3]=y[ik];

                /* initialize statistic variables */
                ravg=0.0;
                aipd=0.0;
                aipd2=0.0;

                /* calculate statistics */
```

```
                    ie=0;
                    for(ic=0;ic < 4; ic++){
                       r[ic]=99999999.9;
                       for(id=0;id < 4; id++){
                        dx=fabs(xsamp[ic]-xsamp[id]);
                        dy=fabs(ysamp[ic]-ysamp[id]);
                        if (id != ic){
                           hypo=xhypot(dx,dy);
                           if (r[ic] > hypo) r[ic]=hypo;
                        }
                        if(id > ic){
                           hypo=xhypot(dx,dy);
                           aipd=aipd+hypo;
                           aipd2=aipd2+pow(hypo,2.0);
                           ie++;
                        }
                       }
                       ravg=ravg+r[ic];
                    }
                    ravg=ravg/4.;
                    aipd=aipd/(float)ie;
                    aipd2=aipd2/(float)ie;

                    printf("%f %f %f\n",ravg,aipd,aipd2);
                 }
              }
           }
        }
     }
}

/* ----------------END OF PROGRAM MAIN---------------------- */

/* ----------------BEGIN UTILITY FUNCTIONS------------------- */

float xhypot(x,y) float x,y;{

   /* FUNCTION: xhypot          WRITTEN: May 1989
      PURPOSE: Function takes two floating point variables
      corresponding to distance between two points on x and y axis
      and returns a floating point variable of the distance
      between the points.
   */

   float z;
   z=pow(x,2.0)+pow(y,2.0);
   z=sqrt(z);
   return z;
}
/* ----------------END OF UTILITY FUNCTIONS---------------- */
```

6.3.3. perm2.c

Important Variables

Variable	Contains
z	in program main, correctin factor for random number generator
x[]	x coordinates of input dataset
y[]	y coordinates of input dataset
COUNT	number of points in input dataset
xsamp[]	x coordinates of sample of 4
ysamp[]	y coordinates of sample of 4
ravg	average nearest neighbor distance
aipd	average interpoint distance
aipd2	average squared interpoint distance

```
#include "stdio.h"
#include "math.h"

    /*  PROGRAM: perm2          WRITTEN: February 23, 1990
        PURPOSE: Program generates nearest neighbor and
        average interpoint distances for 2000 random samples
        of 4 points from the population of all live births
        in tract 5031.04.

        FUNCTIONS: xhypot()
    */

main()
{

    int   ia, ib, ic, id, ie,ig,ih,ij, COUNT,z;
    float x[373], y[373],r[4], hypo,xsamp[4],ysamp[4],
          xhypot(), dx, dy, ravg, aipd, aipd2;

    /*seed random number generator and
      correction factor */
    z=(int)pow(2.,31.)-1.;
    srand(getpid());

    /* initialize array variables */
    for(ia=0;ia < COUNT; ia++){
          x[ia]=0.0;
          y[ia]=0.0;
          if(ia<4){
              xsamp[ia]=0.0;
              ysamp[ia]=0.0;
              r[ia]=0.0;
          }
    }

    /* read in coordinates from standard input */
    for(ib=0;ib<COUNT;ib++){
      scanf("%f %f", &x[ib], &y[ib]);
    }

    /* establish random samples of size 4 */
    for(ig=0;ig<2000;ig++){
      for(ih=0;ih<4;ih++){
            ij=(int)((rand()/(float)z)*((float)COUNT));
            xsamp[ih]=x[ij];
            ysamp[ih]=y[ij];
      }

      /* initialize average variables */
      ravg=0.0;
      aipd=0.0;
      aipd2=0.0;

      ie=0;
```

```
        /* calculate statistics */
        for(ic=0;ic < 4; ic++){
          r[ic]=99999999.9;
          for(id=0;id < 4; id++){
            dx=fabs(xsamp[ic]-xsamp[id]);
            dy=fabs(ysamp[ic]-ysamp[id]);
            if (id != ic){
              hypo=xhypot(dx,dy);
              if (r[ic] > hypo) r[ic]=hypo;
            }
            if(id > ic){
              hypo=xhypot(dx,dy);
              aipd=aipd+hypo;
              aipd2=aipd2+pow(hypo,2.0);
              ie++;
            }
          }
          ravg=ravg+r[ic];
        }
        ravg=ravg/4.;
        aipd=aipd/(float)ie;
        aipd2=aipd2/(float)ie;

        printf("%f %f %f\n",ravg,aipd,aipd2);
    }
}

/* ----------------END OF PROGRAM MAIN---------------------- */

/* ----------------BEGIN UTILITY FUNCTIONS------------------- */

float xhypot(x,y) float x,y;{

    /* FUNCTION: xhypot          WRITTEN: May 1989
       PURPOSE: Function takes two floating point variables
       corresponding to distance between two points on x and y axis
       and returns a floating point variable of the distance
       between the points.
    */

    float z;
    z=pow(x,2.0)+pow(y,2.0);
    z=sqrt(z);
    return z;
}
/* ----------------END OF UTILITY FUNCTIONS---------------- */
```

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
INFORMATION RESOURCES DEPARTMENT
BERKELEY, CALIFORNIA 94720