THE WHOLE PRACTICE AND PHILOSOPHY OF GEOGRAPHY DEPENDS UPON THE DEVELOPMENT OF A CONCEPTUAL FRAMEWORK FOR HANDLING THE DISTRIBUTION OF OBJECTS AND EVENTS IN SPACE.

DAVID HARVEY

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The past decade has witnessed a resurgence of interest in spatial analysis methods by geographers that coincides with the recent interest of geographers in the analysis of longitudinal data using event history models. Much of the recent attention given to spatial analysis methodology may be attributed to the emergence of powerful and graphic-oriented Geographic Information Systems (GIS) technology to desk-top computing (see Fotheringham, 1993; Fotheringham and Rogerson, 1993; Openshaw, 1991; 1992), whilst the current popularity of event history modelling stems from the common finding amongst analysts that it is a more informative and powerful approach to temporal data analysis than traditional methods (see Davies and Pickles, 1985; 1991, Clark, 1992). Interestingly, the models used to analyze event history data are flexible enough to allow spatial data analysis in some cases. The focus of this thesis is the application of event history models to the analysis of a spatial point pattern. The wider goal of this research is to identify and analyze the issues significant to the application of event history models to the analysis of spatial point patterns and processes.

1.2 Spatial Point Pattern Analysis

The description and explanation of the spatial patterns of objects and events are two of the traditional intentions of geographic research (Harvey, 1968). Real world phenomena can be represented on a map by three basic geometric forms: points, lines and areas. Objects or events are represented as points if the sizes of the objects are negligible when compared to the distances between them and the size of the study area. Data in the form of a set of *points* arise in numerous geographical and ecological contexts. An example would be the positions of individual trees in a naturally regenerated forest. In this thesis, any such data set is called a *spatial point pattern* and the points of interest are referred to as *events*.

The reason for the interest in a pattern of events is that, with an appropriate choice of scale, even large objects may be best represented by an event location. It is hoped that by identifying the characteristics of a pattern, some insight into the origins of that pattern will be revealed. Thus, the process acting on individuals, whether they are trees, settlements, stars or other data that can be treated as events, is the main concern of spatial point pattern analysis (Upton and Fingleton, 1985). The statistical analysis of spatial point patterns allows the geographer to obtain a quantitative description of spatial point pattern maps and, in turn, maps of different regions can be compared objectively. In addition, evidence can be gained to help identify the causal mechanisms that underlie the locations of events in space.

Spatial point pattern analysis is distinct from other forms of spatial analysis in that size variations of events are ignored and the matter of interest is basically their existence

on a surface, not the variation in their size (or other characteristics) over space (Upton and Fingleton, 1985). Spatial point pattern analysis methods can summarize and describe map patterns and provide evidence in support of hypotheses concerning the processes which formed the spatial pattern. In contrast, quantitative data techniques such as spatial autocorrelation methods, are applied whenever a variable exhibits a regular pattern over space where its values at a set of locations depend on values of the same variable at other locations (Odland, 1988). Thus, spatial point pattern analysis is useful in contexts where the location, not the value, of an event is significant. Over the past decades a number of methods for the analysis of spatial point patterns have been developed that are now well established in geography as well as other disciplines.

1.3 Event History Analysis

With respect to temporal data analysis, there is an increasing interest in the collection and analysis of longitudinal data or "event histories" (Blossfeld, Hamerle and Mayer, 1989; Tuma and Hannan, 1984). Researchers have recognized that event histories, which record the exact timing and sequencing of events, are often better suited for examining the dynamic nature of social science phenomena than cross-sectional methods. The three primary research areas in the social sciences are unemployment studies, consumer behaviour studies and migration (Uncles, 1988). Examples of phenomena that can be thought of as events include: the births and deaths of individuals or manufacturing plants, the date of leaving school or the completion of a training course, employment changes, promotions, marriage formations and dissolutions, illness,

strikes, crime and many more possibilities. To study such events effectively requires the adoption of an explicitly dynamic approach, and therefore the collection of longitudinal data.

Event history models form a broad class of statistical models based on stochastic processes with discrete states and a continuous time scale. In practical terms, event history analysis involves statistical methods used to analyze time intervals or *durations* between successive state transitions or events. A wide range of models exist, mainly extensions of *survival, reliability* or *lifetime* analysis methods, in which processes exhibit only one event (i.e., one initial and one destination state) as in many biomedical or engineering studies (Cox and Oakes, 1984; Kalbleisch and Prentice, 1980; Lawless, 1982; Tuma and Hannan, 1984). It is precisely these methods that will be considered in this thesis to analyze *spatial durations* between events in space, as opposed to the standard temporal durations.

The literature available in the social sciences on event history analysis has mushroomed in the past decade. Wrigley (1986) has suggested that the "era of longitudinal data" has arrived and many of the studies published in the social sciences share the underlying theme that event history analysis is more informative and powerful than traditional cross-sectional data analysis (Clark, 1992; Davies and Pickles, 1985; 1991; Heckman and Singer, 1986; Tuma and Hannan, 1984). Perhaps this implication will challenge the predominance of cross-sectional data analysis in applied research. More importantly, however, are several extensions to event history models found in the recent literature. For example, Allison (1984) notes that *censoring* and *time-varying*

explanatory variables are common to event history data and are the major impediments to the application of standard statistical techniques, such as multiple regression, to event history data. In fact, the existence of these two problems leads to severe parameter bias and loss of information when analyzed using conventional procedures. However, key innovations, such as the seminal work of Cox (1972; 1975) in regressive event history approaches, has led to rapid progress in the development of theory and method for the analysis of event history data. For example, the partial likelihood estimation method introduced by Cox (1972), is fundamental to partially parametric forms of proportional hazards models that accommodate explanatory variables and censored data. In this thesis, the above developments in regressive approaches for censored observations are exploited in the spatial domain.

1.4 Problem Statement and Research Rationale

Spatial point pattern analysis and event history analysis are similar with respect to the use of frequency distributions of durations separating events. In the case of spatial point patterns, those durations are often distances separating neighbouring events (e.g., settlements). In contrast, temporal durations between successive events are the focus of investigation in event history analysis (e.g., length of unemployment). The basic logic employed in the analysis of temporal durations in event history models has recently been applied to the spatial durations between settlements using *nearest neighbour* to define the duration (Odland and Ellis, 1992). Essentially, this duration based approach is analogous to distance based methods that investigate spatial point patterns based upon the assumed

interdependence between events (see Chapter Two). Examples of such methods include conditions of *complete spatial randomness (CSR)* where event locations are homogeneous and independent, meaning that every location has the same chance of receiving an event and that the chances of a location receiving an event are independent of the locations of other events (Getis and Boots, 1978). Other methods incorporate some form of interdependence among the locations of events such as *contagion* or *inhibition.* These models imply particular frequency distributions for the distances separating events (i.e., the durations) and comparisons of observed frequencies of inter-event distances with the frequencies implied by a model may yield evidence against the operation of a particular process.

As was noted above, the similarity of the distance based methods of spatial point pattern analysis and event history analysis with regards to duration measurements allows for linkages to be made between the temporal and spatial methodologies. This thesis considers the issues of properly defining a *spatial duration*, the inclusion of edge effects with a *censoring* variable and modelling the spatial duration as the response variable depending on a function of explanatory variables. In principle, the range of event history models can be used to control for any type of duration dependence and are explored in this thesis for the analysis of spatial durations. Odland and Ellis (1992) used a proportional hazards model, one class of event history models, to investigate variation in the spacing of settlements in Nebraska. Besides this investigation, the utility of event history methods for spatial analysis has hitherto been neglected by geographers. Consequently, issues regarding the proper definition of a duration, boundary effects and

spatial censoring, and the incorporation of explanatory variables must be investigated further. The methodology in question requires a context; this thesis uses a spatial point pattern derived from the diffusion of an innovation in farming technology in southern Ontario as the data set enabling an investigation of these issues.

1.5 Specific Research Objectives and Significance

The aim of this thesis is to provide additional insight into the utility of event history modelling for the investigation of spatial point patterns. More specifically, exploring the definition of a "duration" in the context of a spatial diffusion process and, subsequently, using these durations as the focus of investigation in an event history modelling framework is the main objective of this research. To achieve this goal, one must consider the data set used in the analysis, the diffusion of an agricultural innovation, as analogous to a typical event history data set where the exact timing of each event (in this case, each adoption) is known. To model the spatial durations however, the location in space of each event, as well as information on a series of explanatory variables for each event (i.e., adopter farm) are used. The coupling of event history methodology and spatial diffusion data provides an opportunity to relate explanatory variables to durations between events and define a "meaningful" spatial duration based on the knowledge from the literature in the substantive field of spatial diffusion of agricultural innovations. The results of this investigation are used to make recommendations for applications of event history models in recovering information regarding the process underlying a spatial point pattern. In theoretical terms, the

currents pulling this thesis are those identified at the onset of this chapter: to further developments in spatial analysis techniques and take advantage of existing analytical methods in the field of event history analysis.

1.6 Outline of Thesis

This chapter has introduced the fields of event history modelling and spatial point pattern analysis; the objective of the study has been identified. Chapter Two will review the techniques, issues and applications of traditional spatial point pattern analysis and, in particular, distance based methods. The chapter extends to a discussion of spatial diffusion of innovation literature with the intention of identifying explanatory variables and an appropriate spatial duration to model. In Chapter Three, a review of the methods of event history analysis relevant to this thesis is presented. Both Chapters Two and Three serve to further justify the rationale and set the context of this thesis. The fourth chapter will outline the data and methods of analysis in detail. The fifth and final chapter will present and interpret the results of the analysis.

THE SHORTEST DISTANCE BETWEEN TWO POINTS IS APPROXIMATELY SEVEN INCHES.

EPHRAIM KETCHALL

EVENT HISTORY MODELLING OF SPATIAL POINT PATTERNS

EVENT HISTORY MODELLING OF SPATIAL POINT PATTERNS: ISSUES REGARDING INTERVAL DEFINITION, CENSORING AND EXPLANATORY VARIABLES

By

PASQUALE ANDREA PELLEGRINI, B.A.(HONS.)

A thesis
Submitted to the School of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree

Master of Science

McMASTER UNIVERSITY

MASTER OF SCIENCE (1993) McMASTER UNIVERSITY
(Geography) Hamilton, Ontario

TITLE: Event History Modelling of Spatial Point Patterns: Issues Regarding Interval definition, Censoring and Explanatory variables

AUTHOR: Pasquale Andrea Pellegrini, B.A. (Hons.) (McMaster University)

SUPERVISOR: Dr. Steven Reader

NUMBER OF PAGES: ix, 178

ABSTRACT

This thesis attempts to further the research by Odland and Ellis (1992) in applying event history models to the analysis of spatial point patterns (i.e., event patterns). Its empirical focus is the event pattern derived from the adoption of an agricultural innovation, the Harvestore, in southern Ontario, Canada from 1963 to 1986.

Event history analysis involves the use of discrete-state, continuous-time stochastic models to investigate a temporal longitudinal record on discrete variables (event history data). Event history models are primarily concerned with durations of time between events and the effects of intertemporal time dependencies on future event occurrences.

Many of the methods used in event history analysis do not preclude the use of other non-negative interval measurements in place of standard temporal intervals to investigate a series of events. In particular, spatial intervals (*durations*) of distances between points (*events*) may also be accommodated by event history models.

This thesis is methodological in nature, and extends the previous research of Odland and Ellis (1992) by using a wider range of parametric models to explore duration dependence, investigating the role of spatial censoring, and using a more extensive set of explanatory variables. In addition, simulation experiments and graphical tests are used to evaluate the empirical event pattern against one generated from Complete Spatial Randomness.

Results indicate that the event pattern formed by the Harvestore adopter farms is clustered (i.e., is described by positive duration dependency). Also, the sales agent is found to be a significant factor in the distribution of Harvestore adopters. In addition, contrasting results obtained from the analysis using censored data versus uncensored data (traditional nearest neighbours) underscores the importance of considering edge effects when using nearest neighbour durations. It appears that an event history approach is a valuable methodology that provides insight into spatial point patterns and processes.

ACKNOWLEDGEMENTS

The completion of this thesis was made possible by a great many people, too numerous to mention individually, who shared their expertise, advice, time and friendship with me over the past two years. In particular, Dr. Steven Reader, my major faculty supervisor, who provided me with an educational experience far beyond the delineation of a Master's degree. I extend my sincere appreciation for his faith in my ability and guidance in moments of uncertainty.

Also, I would like to thank my committee members: Dr. William Anderson and Dr. Pavlos Kanaroglou for stimulating instruction in their graduate courses. I am also indebted to Dr. Barry Boots and Ed Scorgie - without their help this thesis would never have come to fruition. A special thanks to the Geography Office staff: Joan, Darlene, Medy and the others, both past and present, for their gracious assistance and support throughout my undergraduate and Master's degrees.

Many individuals, both in and out of the University environment, were instrumental in providing an intriguing atmosphere during my tenure as a Master's student. Despite not necessarily helping with the completion of this thesis, they did provide more than enough sanity breaks. You know who you are, but for the record, I would like to thank Philip, Jerry and Wendy (my office mates in GS 401) for helping me find those ever-elusive "windows of opportunity." In addition, thanks to the finest group of graduate students one could ask to work side-by-side with, even if they can't play softball. Also, thanks to my very flexible friends who stuck with me in the long run, thanks to Mary, Eric, Phil, Grahme, Nancy, Odette, Franca, Bill and, of course, Lee-Ann. I am most indebted, however, to my family for unfailing support and patience. This thesis is dedicated to my nephew, Matthew, who put everything back into perspective when all seemed lost.

This research was supported, in part, by a Natural Sciences and Engineering Research Council (NSERC) postgraduate scholarship. The financial support of NSERC and the Department of Geography at McMaster University enabled me to pursue my studies on a full-time basis and the generosity of both is gratefully acknowledged.

Pat Pellegrini, June, 1993

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

CHAPTER FOUR: DATA AND RESEARCH METHODOLOGY

CHAPTER FIVE: RESEARCH FINDINGS AND DISCUSSION

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF TABLES

CHAPTER TWO

SPATIAL POINT PATTERN ANALYSIS

2.1 Introduction

The intention of this chapter is to review the relevant literature in the area of spatial point pattern analysis. The chapter begins with a review of the most often used spatial point pattern analysis methods and focuses primarily on "distance based" methods. It is noted that several inherent problems exist when defining distance intervals or durations between events, specifically with nearest neighbour durations. In addition, the problems of boundary events and edge effects are introduced. The next area of consideration centres on spatial diffusion processes from both a geographical and rural sociological perspective. In this regard, the literature reviewed is selected based on its relevance to the empirical investigation of this thesis; that is, diffusion of an agricultural innovation. Here, the literature provides some guidance as to the nature of the spatial durations and explanatory variables that may be examined with event history models. This chapter, along with Chapter Three, set the stage for the conceptual linkages between spatial point pattern analysis and event history modelling pursued empirically in chapters four and five.

2.2 Spatial Point Pattern Analysis

In Chapter One, it was noted that the analysis of spatial distributions and the processes that produce and alter them is a central theme in geographic research. However, researchers in other disciplines (e.g., ecology, biology, geology, forestry, astronomy, and statistics) are also concerned with spatial patterns and processes. As a result, spatial analysis literature is extremely diverse, and unifying these existing approaches has been the goal of a number of texts (Getis and Boots, 1978; Cliff and Ord, 1981; Ripley, 1981; Diggle, 1983; Upton and Fingleton, 1985; Boots and Getis, 1988). In regards to the origins of statistical analysis of point patterns, early studies appeared over 50 years ago in plant ecology literature (Boots and Getis, 1988). During the "Quantitative Revolution" of the 1960s, geographers embraced many of the spatial analysis techniques and have since extended them to facilitate their own research.

The *dot* or *point* map is one of the most common tools used by geographers to display, in a simple and precise manner, the distribution of events in space (Taylor, 1977; Unwin, 1981). The map patterns are assumed as having been created by one or several spatial processes, either human or physical. The analysis of the spatial pattern may reveal underlying causal relationships of events in space. The basic research principle here is that every pattern of events is the result of some process which operated over a region over a time period and this process may be continuing. In this section, static evidence of point patterns is considered, although the intuitive goal is to investigate processes in space and through time.

A *spatial pattern* may be defined as "a zero dimension characteristic of a spatial

arrangement which describes the spacing of a set of objects with respect to one another" (Hudson and Fowler, 1966). A *spatial process* is the force or mechanism generating changes in the spatial arrangement or dispersion of objects, events or groups[1]. A process implies some sequence of reformations which are the result of certain forces (perhaps physical, social, economic or psychological), necessitating a consideration of time, either explicitly or implicitly (Upton and Fingleton, 1985). Thus, the observed spatial pattern is merely a "snapshot" of a spatial process, which may or may not be ongoing, and our goal in analyzing this pattern would be to identify the responsible spatial process(es) which influenced this pattern so that we could predict future "snapshots" or explain similar spatial patterns in other locations. Haining (1990) identifies four important types of spatial processes influencing spatial pattern, namely, *diffusion, exchange, interaction* and *dispersal* (or spread). Section 2.3 defines these processes and concentrates on diffusion processes, with particular attention given to the diffusion of an agricultural innovation.

Popular methods of spatial point pattern analysis proceed by comparing an actual pattern to a theoretical pattern generated from certain assumptions. The most basic hypothesis tested is that of *complete spatial randomness* (CSR), which arises from the *homogeneous planar Poisson point process*. This hypothesis is defined by the following two fundamental properties. First, the number of events in a finite, bounded planar region with area, A, follows a Poisson distribution with mean $\lambda(A)$. In mathematical

[1]More specifically, dispersion refers to the pattern of points with respect to the study area, while arrangement refers to the pattern of points with respect to each other (see Boots and Getis, 1988).

terms, the Poisson distribution can be written as

$$P(n;\lambda) = \frac{\lambda(A)^n e^{\lambda(A)}}{n!} \qquad for\ n = 0,1,2,3,..., \tag{2.2.1}$$

where the parameter λ is called the *intensity* of the process and is defined as the expected number of events per unit area, while n is considered a random variable representing the observed number of events in the region (Figure 2.2.1).

The second property which can be defined as "purely random" or "completely random" (Stoyan et al., 1987) is dependent upon the conditions of *uniformity* and *independence*. Uniformity implies that each location in the region has an equal probability of receiving a event. This means that the study area can be regarded as homogeneous and thus, being completely undifferentiated in space. In addition, this property implies that the spatial point process is the same in all directions from every location regardless of the orientation of A. Hägerstrand (1965) terms the abstract geographical space of the homogeneous planar Poisson point process the "isotropic plane." Finally, independence implies that the placement of one event does not influence the placement of any other event, which means that there is no interaction between events.

Obviously, patterns generated by a homogeneous planar Poisson point process will hardly be observed in geographical reality due to the rigid assumptions it maintains. However, the homogeneous planar Poisson point process can serve as a starting point for models that approximate real world spatial point patterns more closely. Thus, in addition to the random spatial point patterns, two other important general types of spatial point

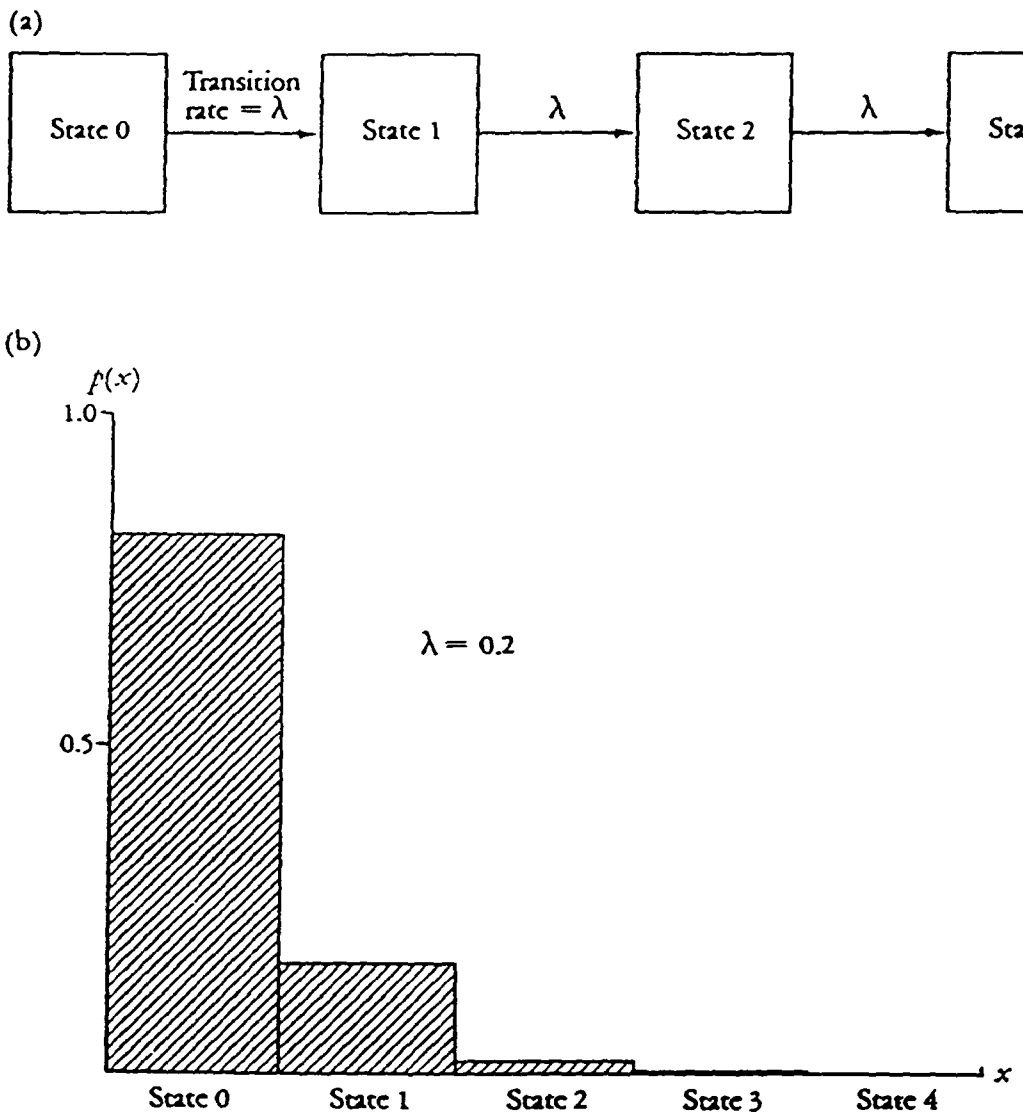

Figure 2.2.1: The Poisson model: (a) the Poisśon process, (b) the Poisson distribution. (Source: Taylor, 1977)

patterns are frequently distinguished. These patterns are termed *regular or dispersed* and *clustered* (Figure 2.2.2). Regular patterns are thought to arise if the assumption of independence is violated in such a way that the events interact by repelling each other. This characteristic of repulsion may then indicate that some sort of competition or inhibition takes place, although environmental heterogeneity cannot be ruled out. Clustered patterns, on the other hand, can be explained by either environmental heterogeneity, which implies that some locations are more likely to receive an event than others, or that groups of events form because events attract each other. Consequently, more information about the study area is needed to decide whether the violation of the uniformity or the independence assumptions has led to the agglomeration of points[1].

The most frequently used techniques to compare the properties of empirical spatial point patterns to theoretical models can broadly be subdivided into two classes: *quadrat analysis* and *distance based or nearest neighbour analysis*. The first class describes the geographic distribution of events according to their density whilst the second describes the geographic distribution of a set of events according to their spacing. The later class of methods is of interest in this thesis. Other methods of spatial point pattern analysis include second-order analysis (Besag, 1978; Diggle, 1980; Getis, 1983; Ripley, 1976, 1977, 1979, 1981), spatial tessellations (Boots, 1981; Boots and Murdoch, 1983; Crain, 1972, 1978; Hinde and Miles, 1980) trend surface analysis (Agterberg, 1984; Chorley and Haggett, 1965), and spectral analysis (Bartlett, 1963, 1964; Tobler, 1969), but are

[1]This is the basic research problem in Odland and Ellis' (1992) application of an event history model to the study of settlement patterns in Nebraska (see section 3.2.5).

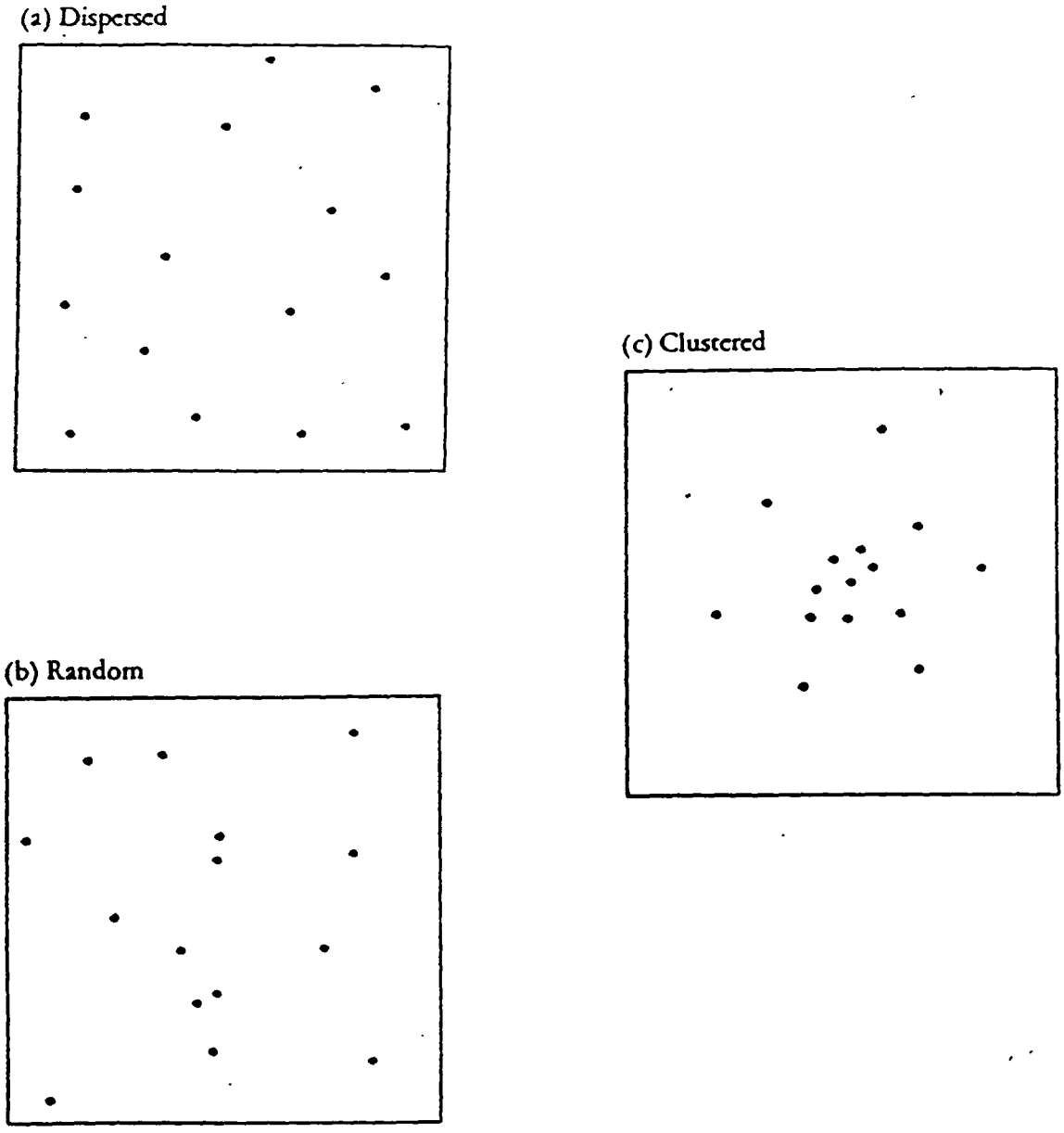

Figure 2.2.2: Three types of spatial point pattern. (Source: Taylor, 1977)

not directly relevant to this research and are not reviewed here.

2.2.1 Distance Based Methods of Spatial Point Pattern Analysis

Distance based analysis, better known as nearest neighbour analysis due to the predominant use of this inter-event measurement, proceeds by comparing the characteristics of the distribution of distances between events and their nearest neighbours in the empirical spatial point pattern with those expected in a theoretical spatial point pattern. This method can be extended to include neighbours of higher order (i.e., second- third-, or higher-order neighbour distances). The general distance based approach is analogous to the temporal duration approach used in event history modelling and hence, is directly relevant to this thesis.

In practice, geographers have used the *z test* and the *R statistic* as descriptive tools for measuring pattern on a two-dimensional surface. The *z* test proceeds by selecting a number of events from the pattern at random. Of course, this means each event has an equal chance of being selected and that the selection of any event in no way influences the selection of other events. This sampling procedure requires the analyst to number each event uniquely and randomly select sample events (Boots and Getis, 1988). The *mean nearest neighbour distance* is then calculated by measuring the nearest neighbour distance, d_i, for each of the sample events, i, and applying the formula:

$$\bar{d} = \sum_{i=1}^{n} d_i / n \qquad (2.2.2)$$

where n is the number of sampled events. Further, the *expected* value of the mean nearest neighbour distance, $E(d)$, for a random sample of events from a CSR pattern is derived from the Clark and Evans (1954) formulation

$$E(d) = 0.5 \sqrt{(A/N)} \; . \tag{2.2.3}$$

Here, the total study area is A, while N represents the total number of events. The observed and expected values may be compared using a normally distributed z statistic. In this situation,

$$z = \frac{[\bar{d} - E(d)]}{\sqrt{var(\bar{d})}} \tag{2.2.4}$$

where

$$var(\bar{d}) = 0.0683 \; A/N^2 \; . \tag{2.2.5}$$

The value of z from tables of the normal distribution is compared to the absolute value of the calculated z as in standard hypothesis tests. The null hypothesis, H_0, is that the pattern under investigation is CSR resulting from a homogeneous planar Poisson process. The alternative or research hypothesis, H_1, is that the pattern is not CSR. In general, when $\bar{d} < E(d)$ and the H_0 is rejected, the implication is that, on average, the individual events are closer than they would be in a CSR pattern. Conversely, if the H_0 is rejected and $\bar{d} > E(d)$, a regular pattern is indicated.

The R statistic is more common in geographical research and is basically the ratio between the actual mean distance and the expected mean distance assuming a random process. That is, the R statistic indexes the event pattern based on the distance between

nearest neighbour events on a surface and may be expressed as

$$R = 2\bar{d}(n/A)^{1/2} \qquad (2.2.6)$$

where all the variables are defined as before. The R statistic tells us how more or less spaced the observed distribution is than a random one on a scale with values ranging from 0 to 2.149 (Figure 2.2.3).

If events cluster on a surface, one on top of another in the same location, so that all durations (or distances between events) are 0, the R statistic is 0. An R statistic value of 0 thus describes a clustered spatial pattern. When $R=1$, the expected and actual average distances between events are equal resulting in a random pattern being indicated. A completely uniform pattern, where events are located as far apart from each other as is physically possible (i.e., a triangular lattice), results in an $R=2.149$ value. In summary, R values of less than 1 indicate distributions tending toward a clustered pattern and R values greater than 1 indicate patterns tending toward dispersion.

2.2.2 Limitations of Distance Based Methods

Both measures discussed above are often used in geographical work, but are subject to a number of obvious limitations. In this section, some of the shortcomings concerning distance based measures such as the nearest neighbour statistic are presented. Generally speaking, investigations which consider only the first-order neighbours do not take into account the direction of each event to its neighbour. In addition, by taking account of only the nearest neighbour distance, information about the pattern is lost. These problems have been the subject of considerable research and second-order analysis

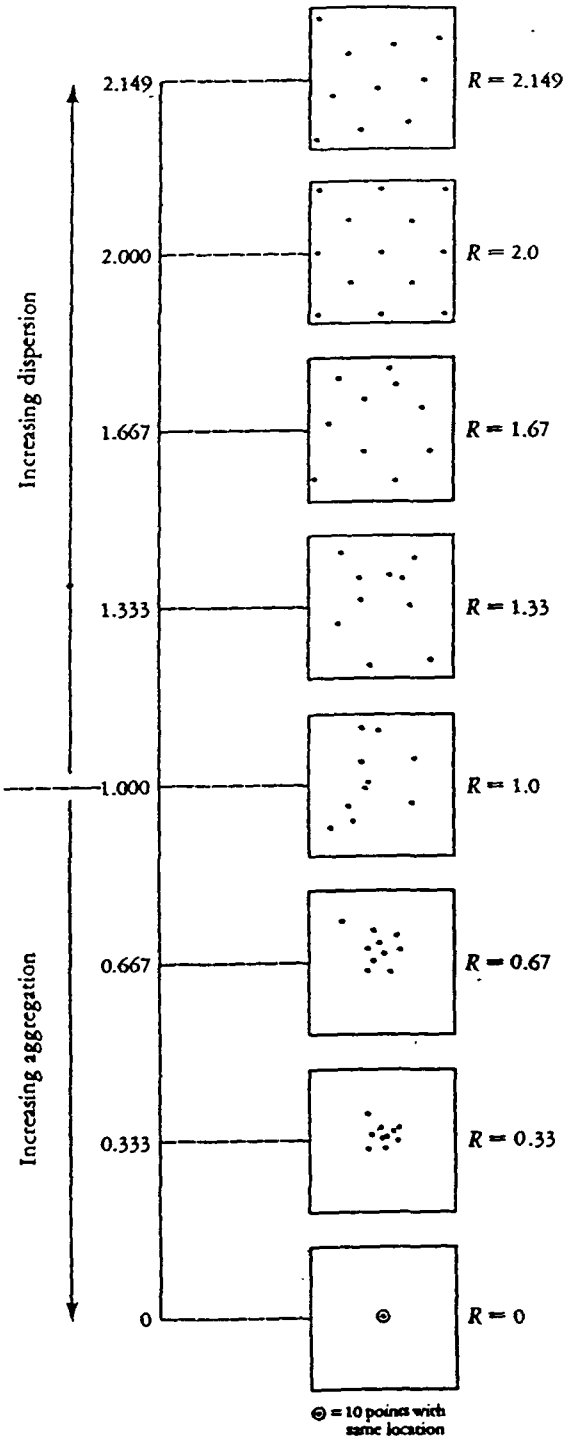

Figure 2.2.3: The R scale. (Source: Taylor, 1977)

and directional statistics are now available in the spatial point pattern literature (see Getis, 1983; Ripley, 1981; Stoyan, Kendall and Mecke, 1987).

Another shortcoming of distance based methods involves the loss of information from the calculation of a single statistic from sets of individual distances measurements. This problem is best illustrated by examining applications of the *R* statistic. Basically, the *R* statistic scale has been criticized in regards to its inability to accurately describe a spatial point pattern. Vincent (1976) and Dawson (1975) illustrate this drawback with the use of special pairs of events, called *dumbbells,* which are reciprocal nearest neighbour events (i.e., they are each others nearest neighbour), whose duration is always smaller than the nearest neighbour distance between any other event in the study region. Figure 2.2.4 illustrates that, by using dumbbells, any number of point patterns can be produced whilst maintaining the same nearest neighbour *R* statistic. In a related issue, the analysts also note the inaccurate use of the terms *regular, random* or *clustered* as they correspond to measures of the *R* scale. In this case, the index describing a random pattern, $R=1$, may also be the index for patterns more similar in form to patterns labelled as clustered or regular.

A more important concern to analysts of event patterns who are constantly working with data in some geographic boundary, is that of *edge effects*. This problem arises in applications of spatial point pattern analysis where the mapped spatial pattern is merely a "window" or study area laid over a real world pattern that extends beyond the mapped area (Figure 2.2.5). In this case, events close to the map border are forced to find neighbours within the mapped area, whereas in the real world the true nearest

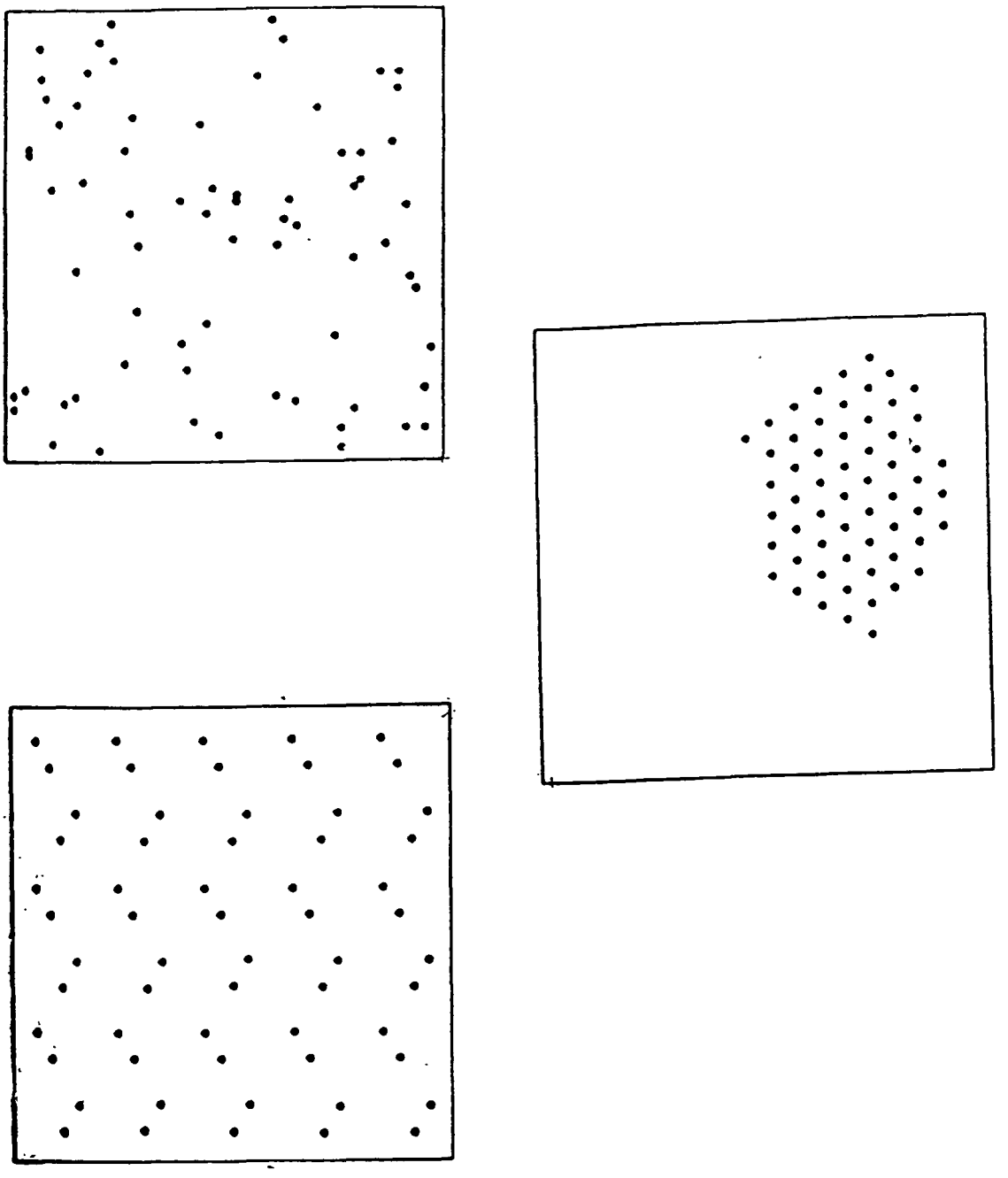

Figure 2.2.4: Three point patterns with an identical nearest neighbour index using dumbbells. Top: random, middle: clustered, bottom: regular (Source: Dawson, 1975)

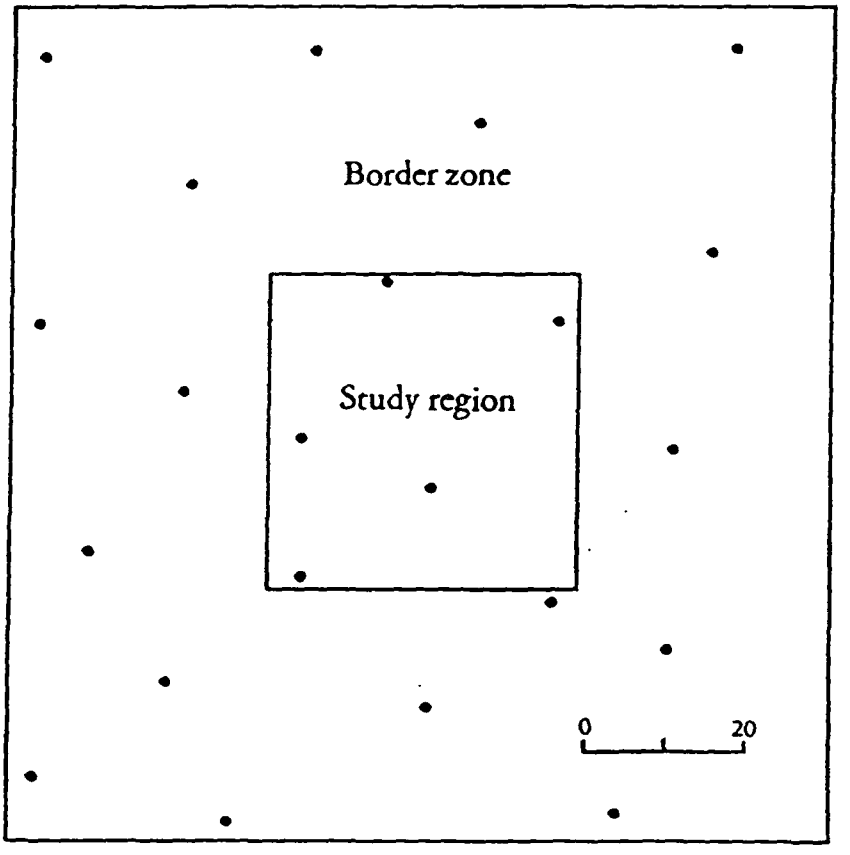

Figure 2.2.5: A spatial point pattern with study region and border zone (Source: Taylor, 1977)

neighbour events may well lie outside the map borders (Unwin, 1981). The main problem here involves the measuring of nearest neighbour durations for events closer to the boundary than to any other event (referred to as *boundary events*). An underlying assumption of Clark and Evan's (1954) derivation of equations (2.2.3) and (2.2.5) is that the values in these equations relate to the nearest neighbour properties of an *infinite* or unbounded CSR pattern. In some situations, applying these equations in the context of a bounded area of a real world pattern can lead to spurious results (Boots and Getis, 1988). For example, nearest neighbour measurements on the boundary events defined above will overestimate the mean nearest neighbour distance.

The problem of edge effects is severe if the number of events under study is small, since the relative number of boundary events tends to increase with decreasing sample size. Solutions to this problem include omission of these events or inclusion of events across the boundary that lie in a predetermined "buffer zone" around the study area. However, to omit boundary events from consideration may decrease an already small sample size, and measuring distances across a boundary may be meaningless (e.g., a pattern of cities on a sea coast) or impossible (e.g., no event data exists outside the study area). A third strategy to compensate for edge effects is known as *toroidal mapping*. If the study area consists of a regular boundary (e.g, a rectangle or square), it may be converted to a torus by joining together the opposite edges of the study area to form a "doughnut" type of diagram (Dacey, 1975; Griffith and Amrhein, 1983). In practice, the study area is then surrounded by identical spatial point patterns and we can assume that the same processes responsible for the location of the events in the original

study area are operating beyond its boundaries (Figure 2.2.6). An irregular study area cannot be handled by the toroidal approach because the boundary edges could not be joined properly without altering the study area.

The fundamental problem in spatial point pattern analysis of boundary or edge effects cannot be ignored in this thesis, just as it must not be ignored in other applications. It was noted that the various parameters estimated in the techniques reviewed in section 2.2 assume that we deal with an infinite plane. In practice, we always deal with a bounded plane. The result is that the processes and patterns defined in theory include no consideration of boundaries. If we are to avoid the bias toward longer nearest neighbour durations for example, that imply a dispersed pattern and are generated by a boundary, then edge effects must be considered. In response to this problem, information on durations to boundaries for all boundary events is incorporated in this thesis as *spatial censoring*, and the value of event history models is that they allow this censored duration information to be included in the spatial point pattern analysis.

The spatial censoring approach departs from the three methods mentioned above for handling edge effects by incorporating the available knowledge on boundary events. For instance, the analyst does have certain knowledge that, for a given boundary event, no other event exists in the spatial pattern up to at least the shortest distance from that event to the boundary. The other methods artificially continue boundary event durations, or require information on out of boundary event locations which may not be available. Of course, the spatial censoring method, by definition, would tend to underestimate the

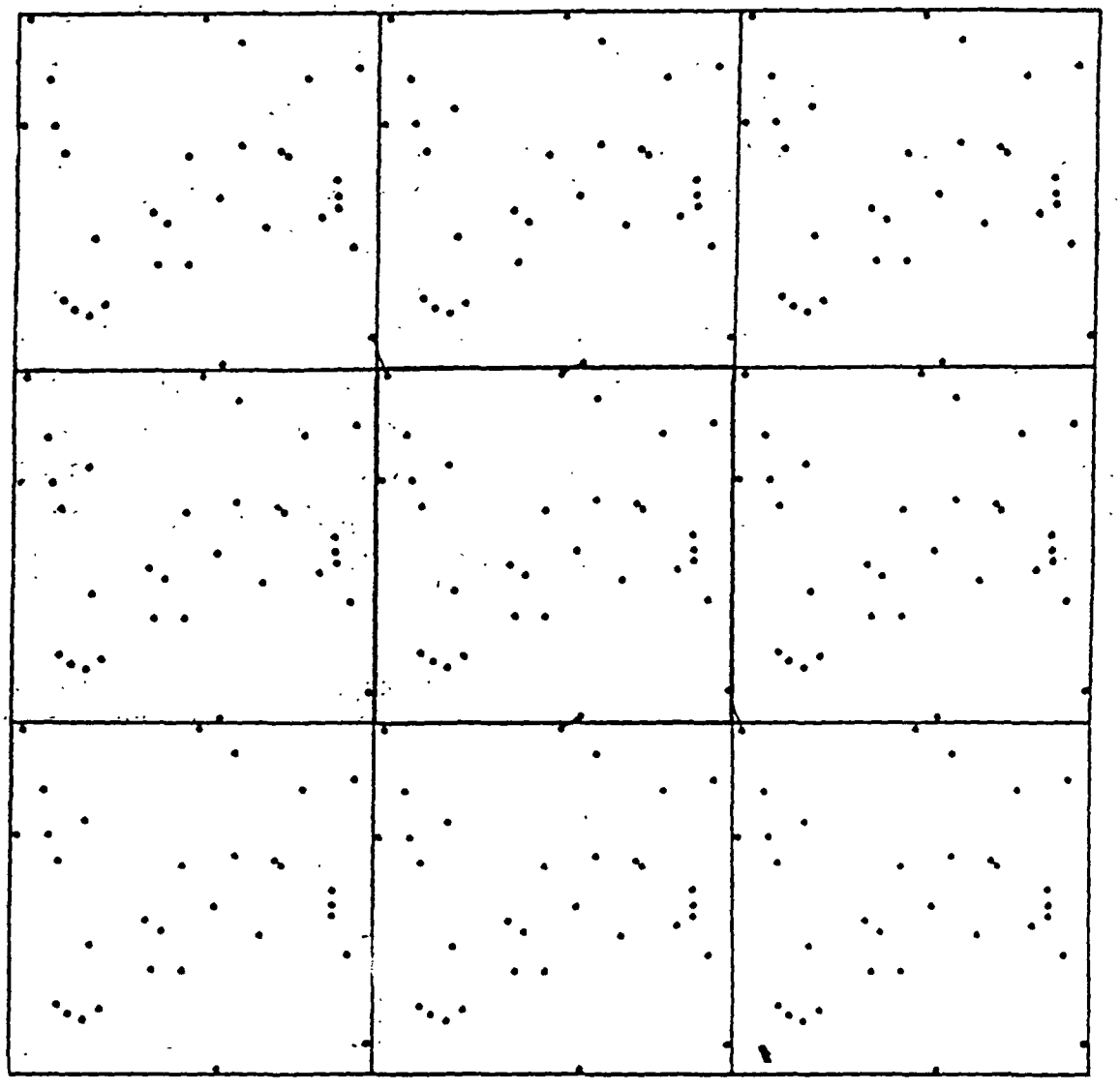

Figure 2.2.6: A toroidal mapping of a study area. (Source: Boots and Getis, 1988)

true inter-event duration. Hence, both a set of durations that include *censored* durations for boundary events (accounting for some edge effects) and an *uncensored* set of durations (where boundary event durations are measured to the nearest neighbour inside the study area) are compared as part of the analysis in this thesis.

Despite the limitations, distance based methods, such as nearest neighbour techniques, have an intuitive appeal as measures of dispersion characteristics of spatial point patterns and, in general, the statistical theory that underlies them is sound. However, in any spatial point pattern there are

$$X = n(n-1)/2 \tag{2.2.7}$$

inter-event distances (or durations) to examine, where X is the number of possible durations and n represents the number of events under examination. The nearest neighbour distances are only one subset of these intervals. This is a general methodological concern for distance based methods, including an event history approach to the study of spatial point patterns. That is, event history methodology, in spatial terms, takes the spatial duration as the basic unit of measurement and hence, the question of defining a duration is a key research issue. In addition, the literature clearly demonstrates the sensitivity of pattern inferences to duration measurements, notably nearest neighbours, and this leads us to consideration of the diffusion of innovation spatial point pattern under investigation to determine which duration to model.

To some extent, the aspects of a spatial diffusion process will guide the process of defining which of the many types of spatial duration should be modelled in order to provide the most meaningful duration. The context of the spatial point pattern is

important here because in the temporal domain, events have a natural order, whilst in space they do not and so it is necessary to impose an ordering on the events to measure a meaningful duration. The spatial diffusion data analyzed in this thesis does, in fact, contain information on the exact timing, as well as location, of each event. Thus, we do have the potential to measure spatial durations using time to order the events (i.e., a nearest-neighbour-in-time duration). However, whether a temporal ordering or some form of spatial ordering is more appropriate to discerning the spatial diffusion process is a legitimate question and is addressed in section 2.3 when the literature on spatial diffusion of innovations is reviewed.

At this point, we should recall that spatial point patterns are only abstractions of reality because, in truth, time never stops and the process(es) generating the spatial point pattern may be ongoing. Whatever the phenomena being studied, processes producing an empirical spatial point pattern are many, and the resulting pattern is much more complicated than the simple limiting case at either end of the R scale. This is evident in empirical studies of settlement patterns in the United States by King (1962) where all three general types of point patterns were found despite the fact that many settlements were service centres and should, theoretically, have shown regularity due to a competitive process hypothesized in central place systems. Furthermore, it should be kept in mind that the spatial point patterns investigated with the aforementioned methods are simplified representations, on the Euclidean plane, of complex objects, where distances are straight lines. Consequently, additional information such as topography, environmental factors, or human decisions that influence the location of geographic

objects are absent.

In essence, no room for causal factors affecting distances between events is allotted. The analyst is left to interpret the process inferred by the particular distance method used based on their substantive knowledge of the subject at hand, but no statistical or regressive analysis is available to aid their analysis. The implication here is that spatial point pattern analysis based on some distance method precludes the use of explanatory variables that potentially provide valuable insights into the process(es) at hand. A simple statistic such as the *R* scale and a classification of the empirical pattern as clustered, regular or random are useful only to limited extent and basically oversimplify the analysis. In this thesis, the event history methodology allows the explicit evaluation of the effects of explanatory variables on duration length. Once again, however, we must turn to the context of the study (spatial diffusion) to provide the necessary explanatory variables, as well as the appropriate duration to model.

2.3 Spatial Diffusion of an Agricultural Innovation

In this section, we review literature concerning spatial diffusion of an agricultural innovation from which the spatial point pattern under investigation in this thesis results. The aim of this section is to reveal, from the spatial diffusion literature, clues for defining spatial durations and explanatory variables to be used in the event history modelling. Diffusion studies, like the spatial point pattern literature discussed above, are also interdisciplinary, but we focus on the work of geographers and their emphasis upon space as a factor affecting the diffusion of innovations.

In an effort to describe and explain rather complex spatio-temporal patterns, geographers have traditionally considered the spatial behaviour of aggregate populations as in urban land use models (Batty, 1976). In fact, some researchers have regarded the spatial behaviour of individuals as both unique and unpredictable (Morrill and Pitts, 1967). However, a wealth of literature in diffusion studies by geographers, sociologists, economists, anthropologists and analysts in the biosciences (e.g., botany and epidemiology) have demonstrated the possibility of focusing geographic research at the level of the individual (see Brown, 1981; Rogers, 1983). Diffusion of agricultural innovations may be viewed as examples of individual level behavioural studies.

In section 2.2 we defined a spatial process and identified four categories of spatial processes influencing spatial pattern (Haining, 1990). We define these processes briefly here, and then concentrate on the spatial diffusion of innovations. First, an *exchange* process is one involving the mutual exchange and commodity transfer between regional, local or national economies. Second, an *interaction* process occurs when events at one location influence and are influenced by events at other locations. Third, mechanisms where the population spreads or disperses and the nature of the spread mechanism influencing spatial structure are termed *dispersal* processes. The final group of processes involve *diffusion* defined as the spread or movement of an attribute or phenomenon over space and through time in a fixed population. Diffusion is distinguished from dispersal since the latter involves the population itself spreading whilst the former entails some variable (e.g., adoption of an innovation) dispersing through a fixed population. For any spatial pattern, the underlying process may be classified into one or more of the above

categories.

In the context of innovation diffusion, a *diffusion process* may be defined as "the process by which an innovation is communicated through certain channels over time among the members of a social system" (Rogers, 1983). The four elements of a diffusion process (innovation, communication, time and the social system) are considered, in turn, below. An *innovation*, in general, is simply "an idea, practice, or object perceived as new by an individual" (Rogers and Shoemaker, 1971). More specifically, technological innovations are "new production inputs, machines, processes and techniques adopted by firms or entrepreneurs for their own use" (Malecki, 1975).

Communication is a process in which participants create and share information with one another in order to reach a mutual understanding and overcome the uncertainty associated with an innovation. It is the newness of the idea that gives potential adopters of the innovation some degree of uncertainty. The uncertainty encountered by an individual can be reduced by obtaining information. According to Hägerstrand (1967a), a basic premise in the conceptualization of the spatial diffusion of innovations is that adoption is primarily the result of a learning process, where an individual adopts an innovation as soon as he/she has accumulated sufficient information to overcome the resistance to adopt. This premise implies that spatial diffusion theory should be concerned with those factors which relate to the spatial pattern of information flow. Thus, fundamental to modelling the spatial aspects of *innovation-adoption* is the manner in which information movement, or communication, from one location to another has been explained (Rogers, 1983).

There are two information sources or *communication channels* identified as being relevant to the learning- and innovation-adoption process of a diffusion of innovation. The first source, *mass media*, is considered important as the initial introduction of an innovation to an individual, but after this introduction creates awareness of the innovation, this source becomes less significant in persuading adoption. The second source, *interpersonal contact* with others who have either (1) previously adopted the innovation or (2) have relevant information and are regarded as reliable sources, is considered more significant in persuading final adoption. For this reason, a central issue in any diffusion study, including this thesis, is the role of the spatial mechanisms of interpersonal contact (Brown, 1981).

The final two elements of a diffusion of innovation process are the involvement of time and the social system. *Time* is an important consideration in the diffusion process since it does not exist independently of events, but it is an aspect of every activity. In fact, most quantitative approaches to the study of innovation diffusion, with the exception of geographic studies, are primarily concerned with time[1]. Time relates to the rate at which the innovation is diffused or the relative speed with which it is adopted by members of the social system. In the present context, the *social system* consists of individuals, organizations, or agencies that share a common "culture" and are potential adopters of the innovation (e.g., farmers, business organizations, residents of a neighbourhood) (Brown, 1981).

---

[1]A comprehensive review of models investigating the temporal diffusion process of an innovation is found in Mahajan and Peterson (1985).

2.3.1 Hägerstrand's Diffusion Model

In spite of the fact that the diffusion of any innovation occurs simultaneously in space and time, research on these two dimensions of diffusion has seldom been integrated (Mahajan and Peterson, 1985). It was noted above that the time dimension has been investigated by researchers representing a wide variety of disciplines, but that spatial diffusion has, for the most part, only been investigated by geographers (Brown, 1981). Hägerstrand's pioneering work in the early 1950s on spatial diffusion of innovations in agriculture provided the initial stimulus for the development of a strong theoretical tradition based on Monte Carlo simulation models (Hägerstrand, 1952, 1965a, 1965b, 1967a, 1967b).

This spatial diffusion work was clearly an attempt to capture, in a diffusion model, the spatial structure of the innovation-adoption process and characteristics of individual behaviour in space. Hägerstrand's approach was deductive and focused on generative processes, and was based on his empirical observation that

> *The spatial order in the adoption of innovations is very often so striking that it is tempting to try to create theoretical models which simulate the process and eventually make certain predictions achievable. (Hägerstrand, 1967b)*

In contrast, much of the other work focusing on diffusion of innovations in agriculture was conducted by rural sociologists and was largely descriptive [for example, Colman (1968), Gross (1949) and Havens (1965) amongst others].

The learning process mentioned at the beginning of this section, wherein information is spread to potential adopters, serves as the starting point for Hägerstrand's conceptualization of spatial diffusion processes. His simulation model focuses

exclusively on the spatial mechanisms of interpersonal contact - the second source of information noted in an individuals learning-adoption process. Briefly, Hägerstrand's methodological framework used Monte Carlo simulation to portray a diffusion process characterized by learning through interpersonal communication - the neighbourhood effect (a contagious process). Adoption takes place after a specified number of messages with information about the innovation are received where the number or messages required varies according to an individual's resistance to adoption. The destination of each message depends on the probability of contact between the teller and potential receiver. This probability is a function of the distance between them. A distance decay function is used to compute the probabilities of interaction between each individual [see Cliff, Haggett, Ord and Versey (1981) for additional details].

The notion of a diffusion of innovation process and the development of a technique for operationalizing this conceptualization are only two of three contributions to present diffusion research credited to Hägerstrand. The third contribution concerns the three empirical regularities he identified during his investigations: the S-shaped adoption curve, the hierarchical effect, and the neighbourhood effect. While Hägerstrand (1952, 1967a) was not the first social scientist to identify these regularities, he is credited with introducing these concepts to the discipline of geography (Brown, 1981; Rogers, 1983). Over time, the cumulative number of acceptors of an innovation is expected to approximate an *S-shape* curve and can be described by the logistic distribution (see Rogers, 1983). Here, the rate of acceptance is low as the innovation is introduced, but as the innovation takes hold in a region, the proportion of adopters quickly rises. The

rate of new acceptances falls off eventually because most people who will adopt the innovation have already done so. In addition, the diffusion is expected to proceed from larger to smaller population centres - the *hierarchy effect*. The final regularity refers to the diffusion proceeding in a wavelike fashion outward from its origin, first progressing to nearby rather than remote locations. This same pattern is expected in diffusion among a rural population and is termed the *neighbourhood* or *contagion effect* (Figure 2.3.7). Hägerstrand (1967a) posits that diffusion in the plane usually displays both neighbourhood and hierarchical spread components (Cliff, Haggett, Ord and Versey, 1981). That is, local communication networks would control diffusion among farmers in a single locale whereas a regional communication network may be operating at the central place level of aggregation (Brown, 1981). A *communication network* consists of interconnected individuals who are linked by patterned flows of information (Rogers, 1983).

2.3.2 Modelling Considerations for Event History Analysis

Since the influential work of Hägerstrand, diffusion research by geographers has shifted towards substantive concerns of the processes under study, rather than the actual mathematical modelling of the diffusion process (see Brown, 1981; Carlstein, 1978; King, 1976; Morrill, 1974). In this section, the essential concepts necessary when investigating a diffusion process are explored to provide the guidance necessary in defining an appropriate spatial duration and the explanatory variables that affect this duration. Based on the important role of communication channels facilitating innovation

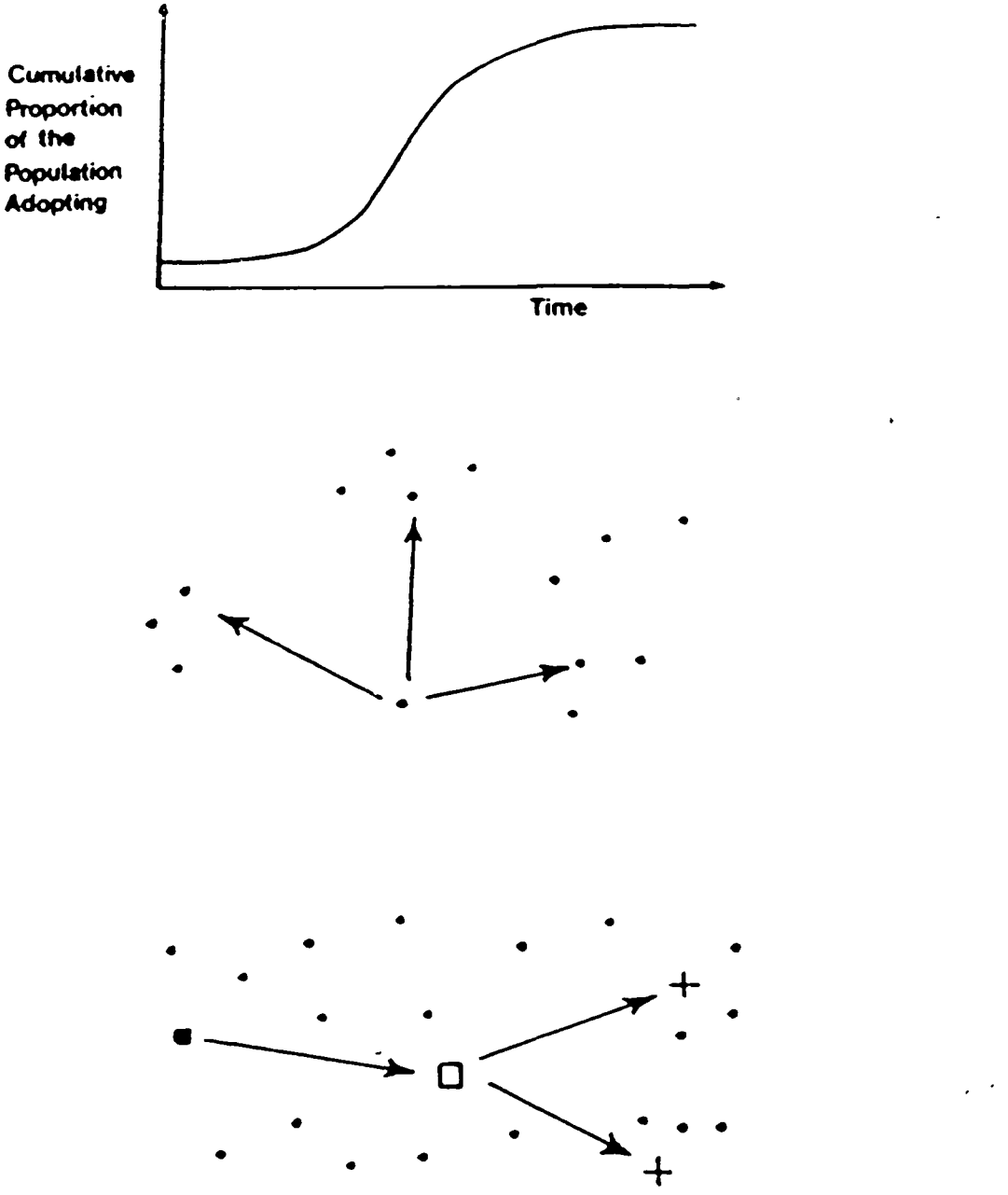

Figure 2.3.7: Empirical regularities in diffusion. Top: S-curve for diffusion through time; middle: the neighbourhood effect for spatial diffusion; bottom: the hierarchy effect for spatial diffusion. (Source: Cliff, Haggett, Ord and Versey, 1981)

diffusion noted in section 2.3, we focus on the micro-scale (individual to individual) **spatial mechanisms of interpersonal contact** as key factors in the spatial diffusion process of an agricultural innovation. The main element in these mechanisms is the neighbourhood effect (or the contagion effect) associated with the communication process in diffusion (Rogers, 1983).

First, we can consider the appropriateness of a temporal ordering of events to define the spatial durations (i.e., the nearest-neighbour-in-time duration). In some ways, it seems obvious that the temporal ordering of events in space can be used to measure spatial durations (Figure 2.3.8). Here, we can imagine some inherent dependence between the location of an adoption event and the previous adoption event in time (i.e., the neighbourhood effect). From a pragmatic modelling perspective, measuring spatial durations as dependent variables in this space-time manner, albeit theoretically attractive, incorporates some systematic bias in spatial duration measurements because of their intimate connection to the scale of analysis. That is, the event density increases with each new event that is added, and hence, the increased occurrence of shorter durations through time over space. Thus, modelling these temporally defined durations will only serve to describe the increasing density of events, rather than the spatial process that determined the dispersion of events.

However, the use of temporal ordering of spatial durations does not explicitly take the spatial mechanisms of information flow and interpersonal contact into consideration. In fact, there is reason to believe that the temporal sequence of adoption may have little to do with the spatial pattern of events (adopters) resulting from the diffusion process.

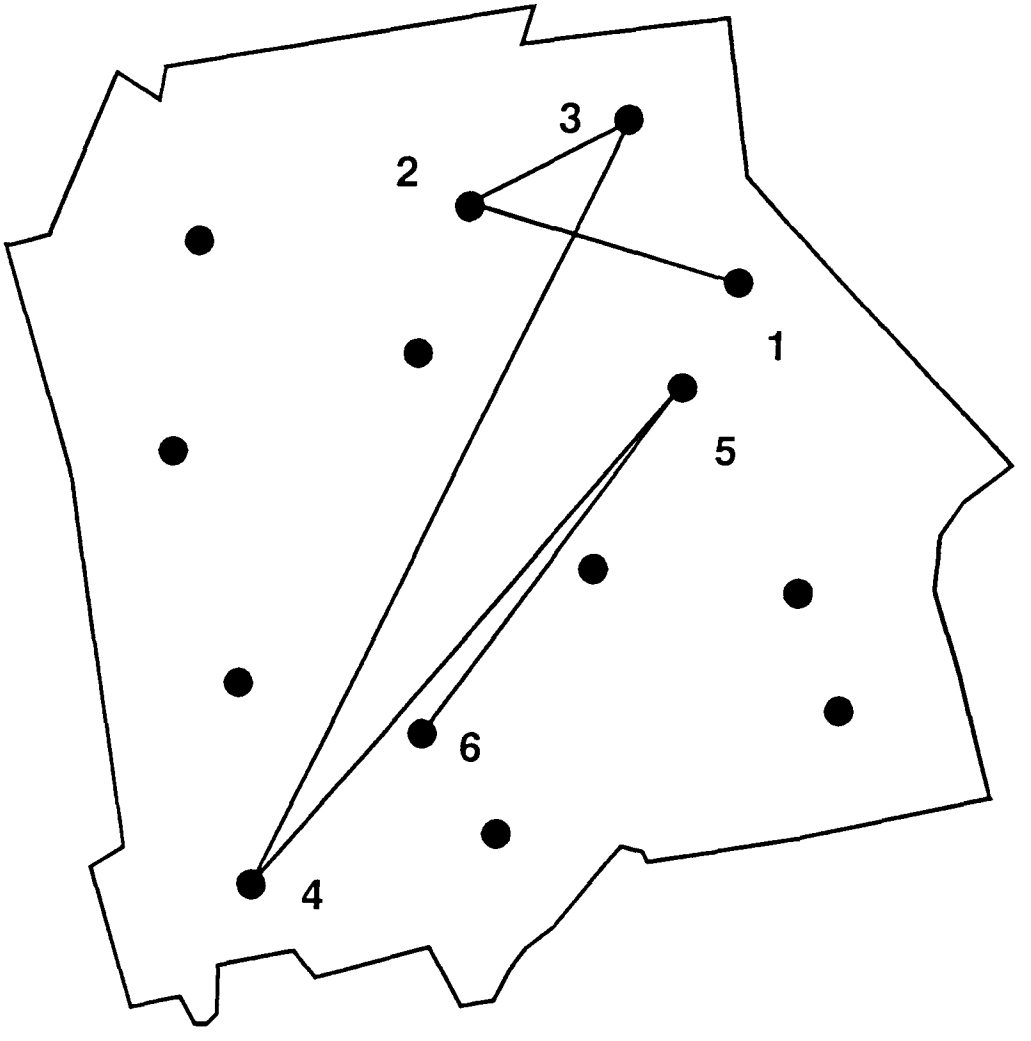

Figure 2.3.8: Nearest-neighbour-in-time duration measurements where the numbering refers to the sequence of events.

For instance, it reveals nothing about the spatial mechanism of communication and spatial contagiousness of the innovation. As an example, consider that an adopter location on the northeastern most region of the study area (event 3, Figure 2.3.8) has little influence, in spatial dependency terms, on an adopter location on the southwestern most region of the study area (event 4, Figure 2.3.8), even if they follow each other sequentially in time. From a geographic perspective, we are not measuring the spatial dependency (how near or far from each other adoption takes place) of the diffusion process directly by following the temporal ordering of events. In fact, one can envision that adopters in close spatial proximity share greater knowledge or communication channels, as opposed to those who adopted the innovation close in time but not in space. Consider, as well, that the communication channels and interpersonal sources of information are different in various spatial locations so that the temporally defined spatial duration may not be the most appropriate to model in order to capture the variation in the spatial diffusion of innovations process.

The alternative to a temporal sequence definition of spatial duration is some form of time independent, inter-event measurement such as the traditional nearest neighbour (Figure 2.3.9). By measuring duration using nearest neighbour (the distance between an event and the nearest other event in the spatial pattern, ignoring time) produces, from a statistical standpoint, independent observations on durations and these observations are readily compared to theoretical standards such as complete spatial randomness (CSR). In addition, we can explicitly model the variation in the spacing of events and any spatial dependencies that may exist. By investigating nearest neighbour durations, each event

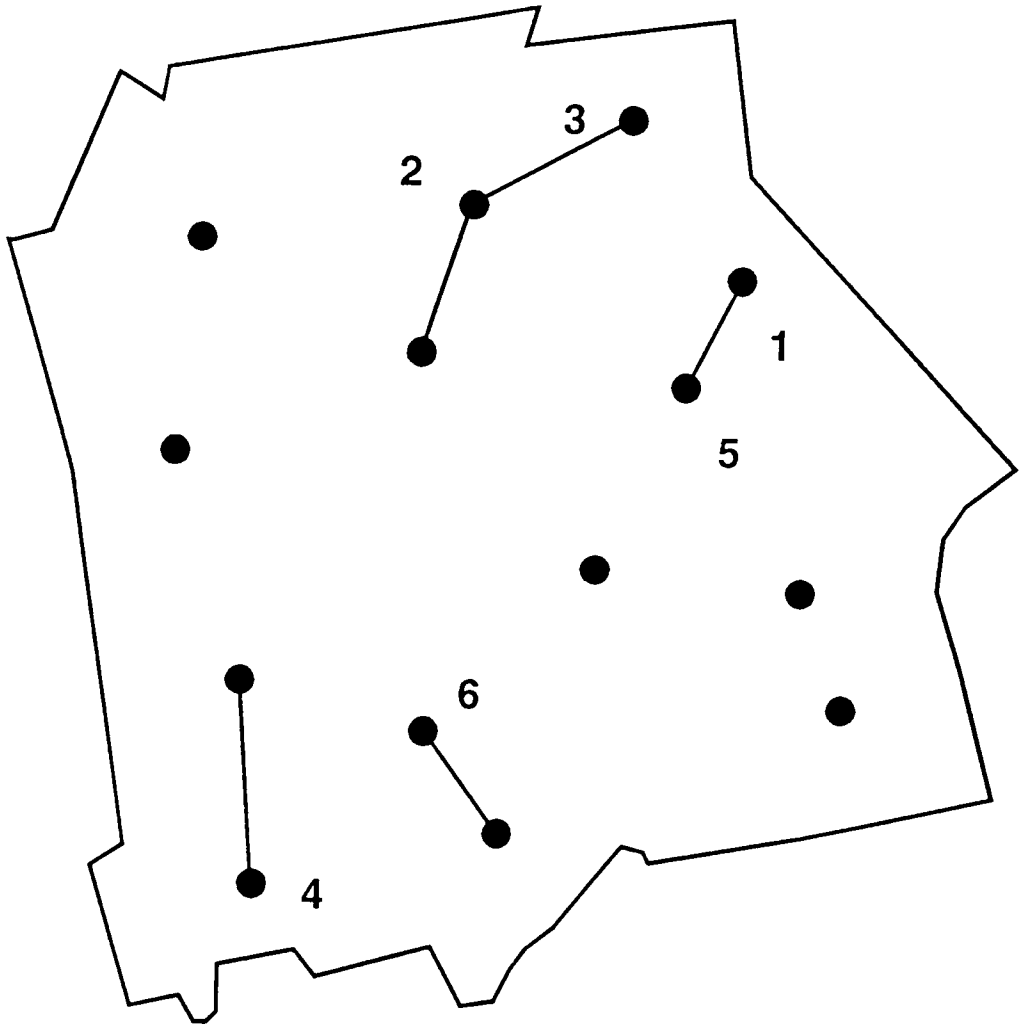

Figure 2.3.9: Traditional nearest neighbour durations where the numbering refers to the sequence of events.

will have an associated duration away from its nearest neighbour as well as a series of characteristics defining this event. The characteristics of each event include sources of interpersonal contact, shared communication channels and regional characteristics (e.g., farming density). In this way, the nearest neighbour duration now maintains some spatial relationship between the distance separating events and the causal mechanisms that may be operating between those events.

This conveniently leads us back to the point made earlier in this chapter on the limitations of distance based methods (section 2.2.2) regarding the absence of a regressive component to nearest neighbour analysis. Given the attractiveness of modelling a nearest neighbour duration, from both a statistical and behavioural representation, we can now focus on incorporating information on the causal factors affecting durations associated with individual events. Event history models allow explicit evaluation of the effects of covariates on given duration lengths, in this case, nearest neighbours between events in space. This provides a distinct advantage over traditional methods which rely on inferences made about the regularity, clustering or randomness of a spatial point pattern to reveal the spatial process(es).

Duration lengths between adopters (i.e., events) may be affected by numerous factors, some easily identified and measurable and others unknown, unmeasurable or both. In terms of known, measurable factors in a spatial diffusion context, interpersonal contacts and communication channels are the keys to innovation diffusion and thus, provide insight into possible explanatory variables. For example, individuals communicating the innovation idea include the change agent, opinion leaders, and

homophilous individuals. A *change agent* is an individual who influences potential adopters' innovation decisions in a manner consistent with a change agency (for example, a salesperson selling a new product). *Opinion leaders* are individuals who are able to influence other individuals' attitudes and opinions based on their informal status in a social system. Also, a fundamental principle of human communication is that the transfer of ideas occurs most frequently between individuals who are alike, or *homophilous*. These similarities include beliefs, education, social status, age or perhaps those in the same farming operation. Interpersonal communication from salespersons, opinion leaders and homophilous individuals are well documented in studies of the diffusion of agricultural innovations (Katz, 1957; Lazarsfeld and Merton, 1964; Lionberger, 1960; Rogers, 1983; Ryan and Gross, 1943), and must be considered in this research to understand the diffusion process operating.

2.4 Conclusions

In this chapter, several important issues have been identified for consideration in the event history methodology being adopted in this research. We introduced traditional nearest neighbour statistics, some of their limitations, and ways of overcoming some of these problems. In particular, it was noted that edge effects are a potential source of bias in any spatial point pattern analysis. The notion of measuring a censored duration on boundary events was then introduced as a mechanism to control for edge effects. Most importantly, the chapter defined the duration to be modelled as the nearest neighbour measurement between events in space. This definition was chosen based on the emphasis

in this thesis on the spatial mechanisms of the diffusion process, wherein events in spatial proximity are thought to be more related than those further apart. Also, the literature on spatial diffusion revealed the types of explanatory variables that could be influencing each event's associated duration. Still to be resolved are the actual explanatory variable definitions, the way in which spatially censored data may be formally included in the event history modelling framework and, more generally, the adaptation of temporally based event history models to the spatial domain. The last two topics are the subject of the next chapter, whilst the defining of explanatory variables is left for Chapter Four where the spatial diffusion data is presented.

WHEN YOU CAN MEASURE WHAT YOU ARE SPEAKING ABOUT AND EXPRESS IT IN NUMBERS, YOU KNOW SOMETHING ABOUT IT.

LORD KELVIN

CHAPTER THREE

EVENT HISTORY MODELLING

3.1 Introduction

This chapter surveys the literature and methods in the field of event history modelling that are used for the analysis of a spatial point patterns in this thesis. The first section looks at the interest of geographers in the field of event history modelling. The second section defines relevant terms and concepts. Next, the three main sources of variation in event history analysis, namely heterogeneity, state dependence and non-stationarity are introduced. The chapter progresses into a summary of the statistical and mathematical concepts of event history analysis used in this thesis. The discussion throughout the chapter focuses on redefining some basic event history concepts in terms of spatial duration analysis, notably the standard parametric and partially parametric model forms and spatial censoring.

3.1.1 Event History Modelling and Geography

The analysis of event histories, or data that follows a given sample of individuals over time, has been the concern of analysts from various fields for several decades. For

example, in the engineering, demography, and biostatistics literature, event history analysis methods appear under the names of "failure-time" or "duration" analysis, "life-table" analysis and "survival" or "lifetime" analysis, respectively (Kalbfleisch and Prentice, 1980; Lawless, 1982). The main interest in event history analysis is the time interval between events, commonly called a duration. For example, in industrial engineering, duration analysis methods are used to describe the useful lives of various machines (i.e., time until component failure) and in the biomedical sciences, to describe events such as the survival times of heart transplant recipients (i.e., time until death).

Many social scientists recognized the natural applicability of such methods to cope with research questions in their respective disciplines. For instance, sociologists have used these methods in life course/cycle research (Mayer and Tuma, 1990) and have made some effort to introduce event history methods to researchers in their discipline (Blossfeld, Hamerle and Mayer, 1989; Tuma and Hannan, 1984). Event history methodology is also familiar to geographers who have had considerable experience with event history data collection and analysis. In fact, Wrigley (1986) notes three historical periods of activity by geographers focusing on event history modelling. The first occurred in the late 1950s when central place systems were studied with manual diary surveys. The second occurred in the early 1970s with geographers using comprehensive diary surveys to study spatial behaviour and urban activity patterns. The third, and ongoing, period of activity concerns the interest of geographers in the methods of analysis rather than the collection of the data. Indeed, an increasing amount of literature describing applications of event history data and methods confirms that this field is well-

known to geographers (Clark, 1992, Crouchley, 1987; Crouchley, Pickles and Davies, 1982; Davies, 1984; 1987; 1988; Davies and Crouchley, 1984; 1985; Davies, Crouchley and Pickles, 1982a; 1982b; Davies and Pickles, 1983; 1985; 1987; 1991; Davies, Pickles and Crouchley, 1982a; 1982b; Halperin, 1985; Pickles, 1983; Pickles, Crouchley and Davies, 1982; Pickles and Davies, 1984; 1985; Pickles, Davies and Crouchley, 1982; Reader, 1992; Reader and Uncles, 1988; Wrigley and Dunn, 1984a-c; 1985). The work of geographers is restricted predominantly to applications of mobility (e.g., social, labour and residential) and consumer behaviour (e.g., store choice), and focuses primarily on temporal dependencies in individual behaviour. One notable exception is the work by Odland and Ellis (1992), which used the partially parametric proportional hazards model of Cox (1972) to examine spatial dependencies in a settlement pattern. The research reported in this thesis attempts to extend the linkage of event history modelling to the spatial domain initiated by Odland and Ellis (1992).

3.2 Terms and Concepts of Event History Modelling

Event History data is defined as data that follows a given sample of individuals over time and, thus, normally provides multiple observations of events associated with each individual. An *event* can be formally defined as a transition between states. In other words, an event may be viewed as a *choice* or *response* that translates as the movement of an individual between *states* (e.g., a product purchase or a residential move) (Allison, 1984). The time elapsed between events is called a *spell* or an *episode*. The length of the observed spells are referred to as *durations* and, typically, the durations

are the dependent variables under study (Kiefer, 1988). Using this terminology, the **events** of interest in this thesis are locations of farms adopting an agricultural innovation, and the **durations** are the Euclidean distances separating nearest neighbour events.

A useful method for introducing concepts of event history data analysis is the five dimensional classification reviewed by Allison (1984). The first dimension, *distributional versus regression methods*, distinguishes event history research which emphasizes describing the distribution of durations from that where the emphasis is on explanation using regressive model forms. More recently, event history analysis methods have focused on regression models in which the occurrence of events is causally dependent on one or more explanatory variables. In spatial duration terms, as with temporal durations, it is useful to look first at which distributional (or parametric) form best suites the duration data prior to adding regressors to the analysis. In this way, the distributional form can describe the general behavioural form of the durations, whilst the regressive analysis provides an assessment of the effects of explanatory variables.

The advent of regressive approaches to event history modelling resulted from the difficulties that arise when more conventional regressive procedures are applied to event history data. Two main problems typical of event history data, *censoring* and *time-varying explanatory variables*, limit the use of standard regressive procedures. To illustrate these problems, the example of recidivism in Allison (1984) is useful. In this study, 430 inmates released from prison were followed for one year after this release. The events of interest were arrests and the aim of the research was to determine how the likelihood of arrest depended on several explanatory variables including race, age at

release, and income. Here, one might be tempted to use the actual duration, or length of time from release to first arrest, as the dependent variable in a multiple regression analysis. However, the analyst is faced with the problem of *event censoring* which refers to the fact that the value of the dependent variable is unknown (or "censored") for persons who were not arrested at all during the one year study period. Of course, if the number of censored cases were small, it might be acceptable to exclude them; but if the data contained a high percentage of censored cases, then exclusion of the censored cases can produce large biases (Tuma and Hannan, 1978). In addition, assigning the maximum time observed, in this case one year, to the censored cases also underestimates the true values and, again, can result in considerable bias.

Beyond the problem of censored observations when modelling durations, the analyst must now face the task of incorporating explanatory variables that change in value over the observation period. These variables are termed *time-varying explanatory variables*. To further complicate matters, time-varying variables can also vary across the observation period, yet be static within certain durations (in a multiple spell situation, see the discussion below). Using the same example of recidivism data from above, it is reasonable that the variable *income*, for instance, may change over the one year period of study, and any changes in income may be incorporated by conducting follow up interviews with the individuals, perhaps on a monthly basis, to update income data. Although including twelve income measures (i.e, one for each month) seems reasonable, albiet long winded, for the person arrested in the twelfth month, it is inappropriate for persons arrested during the first few months after release. In this regard, the person

arrested after one month may have been incarcerated during the remainder of the observation period, so his/her income becomes a consequence, rather than a cause, of recidivism. In essence, his/her income after the month of arrest is irrelevant to tha analysis.

Both censoring and time-varying explanatory variables are common to event history data, but, fortunately, several innovative approaches to handle these problems have been developed by the interdisciplinary researchers using event history data and methods. The notion of event censoring, and its role in the presence of potentially severe edge effects in a spatial context, is explored further in section 3.2.2. In terms of a spatial analogue to time-varying variables, we look at spatially varying variables at the end of the chapter (section 3.2.5) when the discussion turns to Odland and Ellis' (1992) analysis of variations in the spacing of settlements with an event history model. In this thesis, event censoring is incorporated into the analysis, but spatially varying explanatory variables are not included.

Allison's (1984) second dimension of event history modelling, *repeated versus non-repeated events*, distinguishes the work within biostatistics and engineering and that of many geographic and social science applications. In this regard, many social scientists are concerned with events that may occur numerous times to each individual as in residential mobility or employment status studies. In contrast, much of the biostatistical work, as in the example of heart transplant recipients at the beginning of the chapter, is concerned with single (or *terminal*) events. In this case, the transition in state (or event) is death which is a final, absorbing state. This thesis is concerned with survival type

analysis (i.e., single duration data) within the broader field of event history analysis. In Figure 3.2.1, we see that all the spatial durations are converted so each duration is unidirectional, has the same "starting point," and each event has only one duration associated with it[1].

The third dimension, *single versus multiple kinds of events*, refers to the analysis of durations measured for various event types. In the field of biostatistics, for example, it is important to distinguish deaths due to cancer from deaths due to other causes, and analyze the duration data accordingly. Formulations known as *competing risks* models accommodate multiple event type data and have been used in studies of housing careers in the geographic literature (see Pickles, 1983; Pickles and Davies, 1985). It is quite likely that multiple event type models will appear more frequently in social science applications given that multiple event type, multiple duration data are the norm. In this thesis, only one event type, the adoption of a farming innovation, is considered, so competing risks formulations are not explored.

The fourth dimension, *fully parametric (or partially parametric) versus nonparametric* methods, concerns the treatment given to the distribution of duration lengths. In the social sciences, the typical fully parametric approach is to specify a particular distributional form for the durations such as the Weibull or exponential distributions. On the other hand, nonparametric methods make no assumptions about the distribution of durations. The *proportional hazards model*, developed by Cox (1972), is a *partially parametric* (or semi-parametric) model in that it does not specify the exact

---

[1]In this respect, the spatial durations now resemble temporal cohort data.

Figure 3.2.1: Converted nearest neighbour durations.

form of the distribution of durations, whilst a functional form is specified for a regressive component in the model. This partially parametric model is common in the biomedical literature (Cox and Oakes, 1984), and has recently been applied to spatial durations by Odland and Ellis (1992). In this thesis, both fully and partially parametric modelling approaches are used to examine spatial durations separating the locations of innovation adopters.

The final dimension identified by Allison (1984) is the *discrete time versus continuous time* methods of analysis for event history data. This dimension can be thought of as "discrete space versus continuous space" methods for the purposes of this thesis. Essentially, methods that assume the measurement of durations are exact would be classified as continuous. In contrast, when the measurement units of durations are large enough that they do not permit very precise measurements, they are termed discrete. To some extent, the distinction is relative. For instance, when the discrete units are very small, it is usually acceptable to treat the duration measurements as if they were, in fact, on a continuous measurement scale. The spatial event data used in this thesis is continuous since event coordinate locations may, in theory, be measured to any level of precision, and hence, continuous time event history models are used.

3.2.1 Spatial versus Temporal Variation

Models containing stochastic as well as structural components have been constructed in an attempt to understand three inter-related sources of behavioural variation in event history modelling. The variation amongst a population of events in

space is very different than variation in time, but some analogies may be drawn from the temporal domain that are useful for understanding spatial processes. First, we should clearly state the distinction between a temporal and spatial series of events. In the case of a series of events randomly distributed in time, for example, the proliferation of cells in an organism, the statistical methods used for analyses of these processes reflect the essentially unidirectional quality of the time dimension. Hence, an investigation for a temporal series of events might depend on the inherent ordering of the observations from earliest to latest. In contrast, to investigate a series of events in space, no corresponding directionality exists; hence, the problem of defining which spatial duration to measure and, subsequently, model. In other words, the analysis must impose some "ordering" or dependency structure on the spatial events. In the following discussion, the sources of variation that can be distinguished by event history analysis are defined in temporal terms for clarity of presentation, but subsequently redefined in spatial terms for this thesis.

Typically, event history data for a group of individuals allows analysts to model the probability of a change in state, or the occurence of an event, as some function of three sources of behavioural variation. If we take as an example a group of shoppers and use event history analysis to study the effects of grocery shopping interpurchase times on switching shopping centre choice, then the *durations* are the time intervals between grocery shopping trips, and the *events* are the store switch (Reader and Uncles, 1988). To understand the individual consumers behaviour and appropriately model the probability of a store switch, three main sources of variation, namely *heterogeneity, state*

dependence and *non-stationarity*, are incorporated into the model. Heterogeneity is defined as those measured variables which in some way influence shopping behaviour, such as income, car ownership and other socio-demographic or interpersonal variables. Since these variables are known and measurable by the analyst, this variation is more suitably termed *observed heterogeneity*. Conversely, *unobserved heterogeneity* comprises those variables which might influence behaviour, but have not been measured, either because they have not been collected or because the variables are unobservable (e.g., individual motivation or tastes) (Reader, 1988). *Non-stationarity* refers to the variation over time of individual event probabilities due to the effects of time-varying exogenous variables (related to the process environment or the individual). In other words, non-stationarity may include the effects of store promotions (process environment) or an income change (individual) on store switching for our shopping example.

The third major source of behavioral variation, *state dependence*, is a general term incorporating the effect on behaviour of the individual consumer's record of events. The consumer's record of events, or the probability of a change in state, is related to one of four different forms:

(1) *Markovian*: the current state, or previously occupied states for higher order Markovian effects.

(2) *Occurrence*: the number of times different states have been occupied.

(3) *Duration*: the length of time the current state has been occupied.

(4) *Lagged Duration*: the lengths of time previous states have been occupied where all previous states, or only those durations occurring in the same state under investigation,

may be considered (after Heckman and Borjas, 1980).

The main focus of research concerning state dependence is the need to distinguish true state dependence effects from spurious state dependence effects to prevent biased parameter estimates (Davies and Crouchley, 1984; Pickles and Davies, 1984). *True state dependence* refers to the influence of past behaviour on future choices based on real barriers or aids to a change in state, such as benefits (utilities), commitments or motives. True state dependence is also called *structural dependence, feedback effects, cumulative inertia and true contagion. Apparent state dependence*, in contrast, arises when individuals differ in their propensity to experience an event (i.e., heterogeneity). When these differences are correlated over time, previous experience will appear to determine future events or choices.

Several well-established examples of the apparent state dependence effect that introduces bias into the parameter estimates of the observed exogenous variables are available from the literature in the social sciences. In the context of unemployment durations, apparent state dependence may arise in the presence of an omitted variable, for instance, the "motivation" of individuals to seek new employment. This variable is independent of the observed exogenous variables, but is positively correlated with the chances of obtaining work. The observed event history data will show a decline in the chances of obtaining work with increasing duration of unemployment producing an apparent state dependence effect and misleading results (Davies and Pickles, 1985a, 1985b). Therefore, apparent state dependence is not a result of any inter-temporal state dependence, rather it is an indicator of unobserved or unmeasurable variables that persist

through time. *Spurious state dependence* and *spurious contagion* are equivalent terms for the related phenomena of omitted variables and apparent state dependence (Davies and Crouchley, 1984; Davies, Pickles and Crouchley, 1982b; Davies and Pickles, 1983; Pickles and Davies, 1984, 1985; Pickles, Crouchley and Davies, 1982).

Translating the temporal sources of variation to their spatial analogues is essential so that spatial sources of variation may be considered in the event history analysis of spatial durations. Here, heterogeneity and non-stationarity are more difficult to discern in spatial terms. In fact, Odland and Ellis' (1992) use of an event history model to examine spatial heterogeneity in a settlement point pattern is actually, in spatial point pattern terminology, an attempt to measure spatial non-stationarity in a region. Perhaps, then, a more general term is appropriate, for example, *environmental heterogeneity*, which describes the variation in the propensity of different locations in a region to experience an event. In Chapter Two, we noted that clustered patterns may be the result of either some spatially contagious process or, alternatively, environmental heterogeneity where the event clusters represent locations that are more likely to contain an event than others. In this thesis, we incorporate explanatory variables to account for some of the environmental heterogeneity in the spacing of adopter events. However, this term is not restricted to process environment variables, such as farm density measurements across a rural landscape as in an agricultural diffusion study; but also incorporates event-specific characteristics that might influence the spacing of neighbouring events. For example, various characteristics of farmers (e.g., farm operation type, feed type preference) that may influence the duration associated with those adopter locations, and should be

included in a regressive analysis.

State dependency, unlike heterogeneity and non-stationarity, is more easily reinterpreted in spatial terms. In this case, *spatial state dependency* involves endogenous factors produced by, or directly related to, the spatial process. The same four types of dependency from time can be identified in spatial terms as the probability of an event occurring at a location could be modelled as dependent on: the nearest current event (*Markovian*), the number of events (*occurrence*) within a specified distance, the length of the nearest neighbour duration (*duration*) or several nearest neighbour durations (*lagged duration*). In this thesis, event history hazard rate models with and without **duration dependency** are used to analyse spatial dependencies between adopter farm locations. Given the preceeding discussion, we must be aware of the potential for spurious duration dependency arising from unobserved spatial (or environmental) heterogeneity. For example, some events may occur at large nearest neighbour durations from one another. This may be the result of a situation where areas in space that are predisposed to an event occurrence are saturated, and so a greater share of events occur at distant locations. The longer spatial durations associated with this phenomena may mislead analysts if the variable for predisposed areas (i.e., unobserved heterogeneity) is absent from the model.

3.2.2 Event Censoring and Edge Effects

The notion of *event censoring* was presented earlier, but is explored more fully here. As with the previous section's discussion, event censoring is first defined in traditional temporal terms, and then redefined in a spatial context. In general, event censoring occurs when the time until the first event (left-censoring) or the time until the event following the end of the observation period (right-censoring) is not fully known (Kiefer, 1988). The loss of information stemming from either meaningful (i.e., birth date for cohorts) or arbitrary (i.e., survey inception) starting dates is termed an *initial conditions* problem and is related to the *left-censoring* problem. This problem usually falls under the unobserved heterogeneity component from a modelling standpoint since it refers to omitted or unmeasurable variables connected to the time period before the survey.

If event occurrence probabilities are assumed dependent on the time since the last event, then naturally left-censoring is a concern. The degree of parameter bias due to censoring varies with the model being estimated, but as the duration of the study period increases relative to the mean duration times between events, the bias decreases (Flinn and Heckman, 1982). In contrast, failure to account for left-censoring especially in *low-involvement* situations (i.e., where the number of events per individual is small and thus the influence of any lost information is important) results in significant parameter bias (Davies and Crouchley, 1984; Tuma and Hannan, 1984).

Left-censoring may be considered problematic since the censored observation has a relationship to the past history of the process on which no information exists. In

contrast, the problem of *right-censoring* or *final conditions* relates to lost post-sample information and, although frequently encountered, is less problematic since the censored observation is simply an unobserved continuation of the observed process (Tuma and Hannan, 1978). Ignoring the available information on right censored observations would underestimate the true *survival rate* or the probability that a duration extends to at least some specified time. Again, final conditions may arise from meaningful (i.e., a terminal event such as component failure or death) or arbitrary (i.e., termination of survey) event dates, similar to initial conditions.

Figure 3.2.2 illustrates a typical event history data set which represents a complete marital history including state occupancies and times of events. The horizontal axis in Figure 3.2.2 represents time and the vertical axis shows the individuals status at a given point t with regard to the three states: dead, married, not-married. The dashed vertical lines denoted τ_1 and τ_2, identify the observation period for these individuals. The information to the left and right of this set is an example of censored data, since their values are unknown to the researcher. Kalbfleisch and Prentice (1980) note that the necessity of obtaining methods of analysis that accommodate censoring is probably the most important reason for developing specialized models and procedures for event history data.

In this thesis, **spatial censoring** is defined as the shortest distance from a boundary event to the boundary. This distance is measured by extending a line from the boundary event (i.e., adopter location) to a boundary point so that the line is perpendicular to the boundary. Recall that boundary events are those events whose

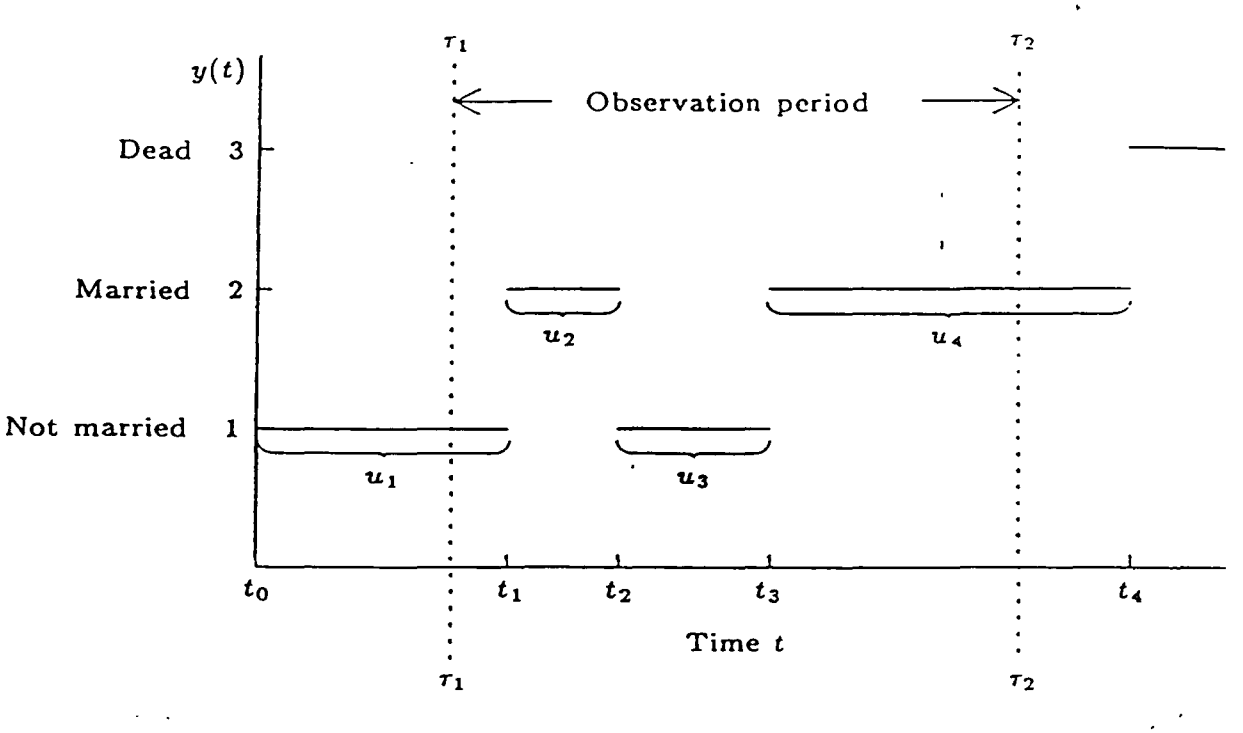

Figure 3.2.2: Hypothetical marital history of a typical person. (Source: Tuma and Hannan, 1984)

distance to the boundary is less than their associated nearest neighbour duration. Spatial censoring is related to "right-censoring" in temporal terms since the censored observation is an unobserved continuation of an observed process. In other words, we use data from events where the duration has not "terminated" at a nearest neighbour duration, and so is actually incomplete (Figure 3.2.3). This means that the analyst has certain knowledge that no other event, up to the censored duration measurement, is closer to that boundary event, and this information can be included in model parameter estimation. Notice, however, that the true length of the duration is actually unknown, but that forcing boundary events to find a nearest neighbour would tend to overestimate the nearest neighbour durations (see Chapter Two). A first step at examining the utility of spatial censoring is to compare the analysis results of the data set with some censored durations to the conventional nearest neighbour durations (i.e., *uncensored* durations) results, where both sets of observed durations stem from the same empirical event pattern. Thus, both an uncensored and censored set of spatial durations are analyzed in this thesis. For the purposes of this thesis then, the censored duration measurement provides absolute knowledge of the distance around a boundary event where no other event exists.

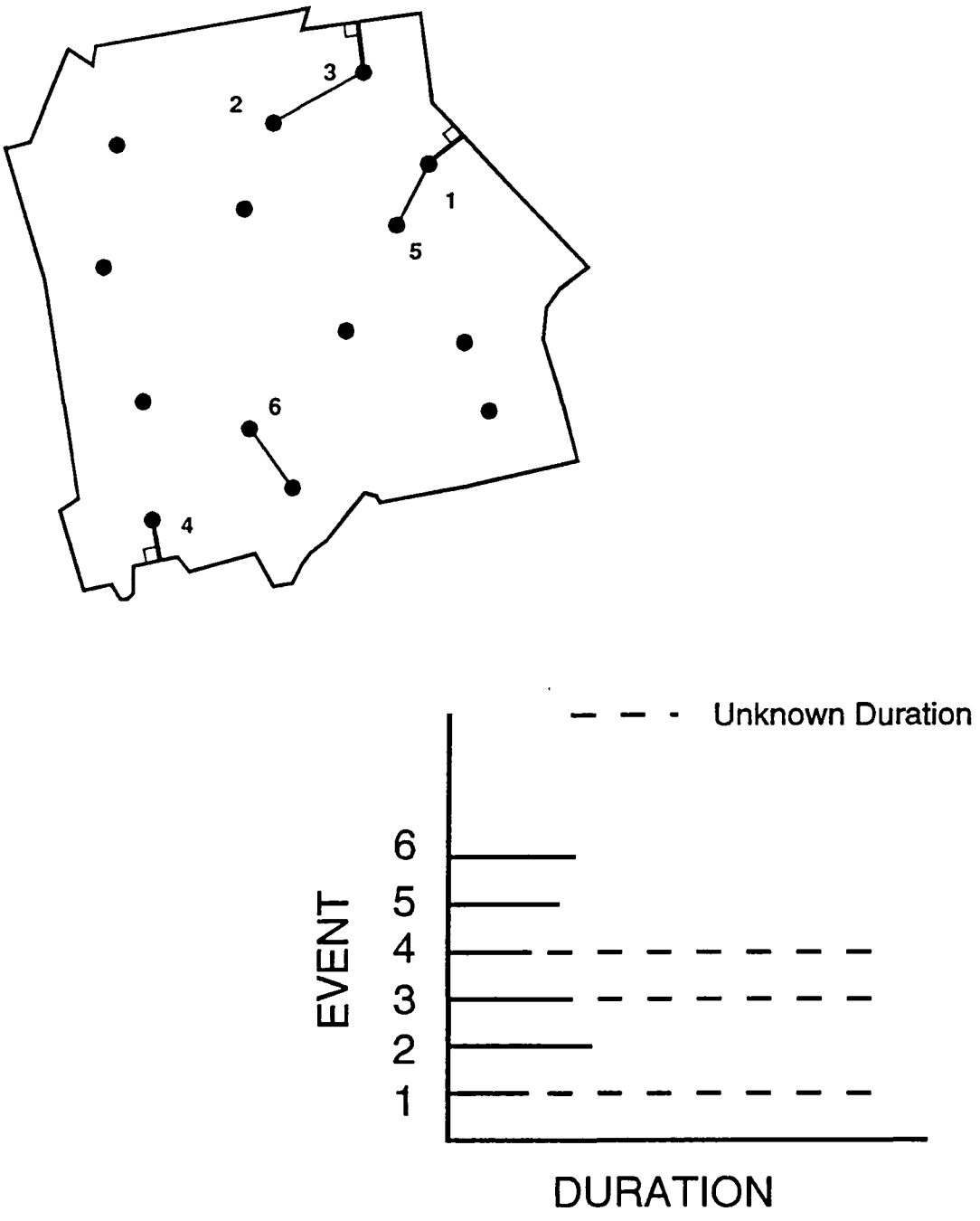

Figure 3.2.3: Spatial durations with some censored duration measurements.

3.2.3 Statistical Concepts of Event History Analysis of Spatial Durations

Although developed for the analysis of temporal durations, the formulations of event history models do not preclude the use of other non-negative random variables, for example, values of distance. The discussion that follows applies to *spatial durations* and is confined to models where population homogeneity is initially assumed, whilst explanatory variables are introduced in section 3.2.4. Here, space is a continuous variable since an event (or adoption of the innovation) may occur at any location. The length of the distance measurements between nearest neighbour events define a spatial duration - the basic observation in these models.

The class of statistical models known as survival or duration models, within the broader field of event history modelling, have generally been developed on the basis of the *probability density, survivor* and *hazard* functions. The *probability density* and *distribution* functions for the duration $X(X \geq 0)$, where X is a non-negative, continuous random variable, are denoted by $f(x)$ and $F(x)$, respectively. In mathematical terms,

$$F(x) = P(X \leq x) = \int_0^x f(u)\,du, \tag{3.2.1}$$

and for all points for which $F(x)$ may be differentiated

$$f(x) = F'(x). \tag{3.2.2}$$

The *survivor* function is

$$S(x) = P(X \geq x) \tag{3.2.3}$$

and expresses the probability that the duration extends (or "survives") until distance x. Alternatively, the survivor function can be expressed as

$$S(x) = 1 - F(x) \tag{3.2.4}$$

where $S(x)$ is a non-increasing function of distance approaching zero as distance (or duration) increases (see Figure 3.2.4). The probability density function may be equivalently expressed as

$$f(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x)}{\Delta x} . \tag{3.2.5}$$

The *hazard* rate (also known as the *hazard function, failure rate,* or *risk* function) is defined as the instantaneous probability that durations in the interval x to $x+\Delta x$ end, given that they extend at least to x (Tuma and Hannan, 1984). The hazard rate is normally interpreted as the rate at which durations of time come to an end, but, in this thesis, the durations are values of nearest neighbour distances. The hazard rate is specified as

$$\lambda(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x \mid X \geq x)}{\Delta x} . \tag{3.2.6}$$

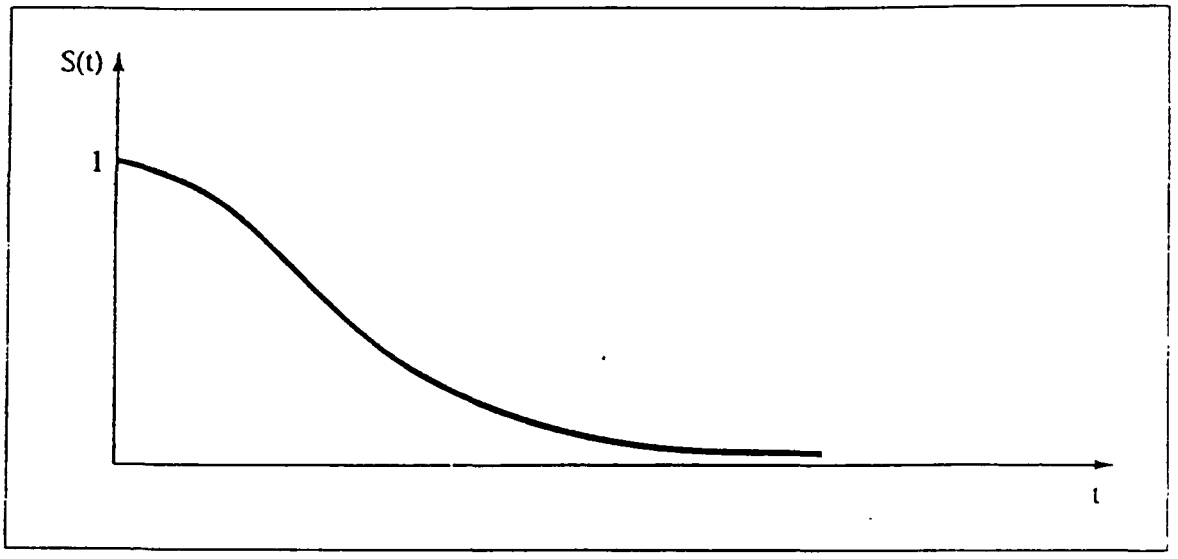

Figure 3.2.4: Typical shape of a survivor function, where t=x. (Source: Blossfeld, Hamerle and Mayer, 1989)

The hazard rate defined above and formulated in equation (3.2.6) is the key concept used in the analysis of event history data. In temporal analysis, the hazard rate contains information concerning the future course of an individual under study, given that they have survived (i.e., no event has taken place) to a certain time, as opposed to a distance, x. In most cases, researchers have some *a priori* information concerning the form of the hazard rate from the substantive theory of the current application. An example from temporal analysis is the hazard rate for human mortality which is characterized by the *U* or *bath-tub* shaped curve (Figure 3.2.5). The curve is explained by the following process: due to high infant mortality at the beginning of the life cycle, the hazard rate of dying is high, then decreases and remains approximately constant until aging causes it to rise once again (Blossfeld, Hamerle and Mayer, 1989; Kalbfleisch and Prentice, 1980). Monotonic increasing, monotonic decreasing or constant hazard functions are also found in various substantive contexts.

The relationship between the hazard rate and the survivor function is derived from equation (3.2.6) as

$$\lambda(x) = \frac{f(x)}{S(x)}, \tag{3.2.7}$$

substituting equation (3.2.4) we have

$$\lambda(x) = \frac{f(x)}{1 - F(x)}. \tag{3.2.8}$$

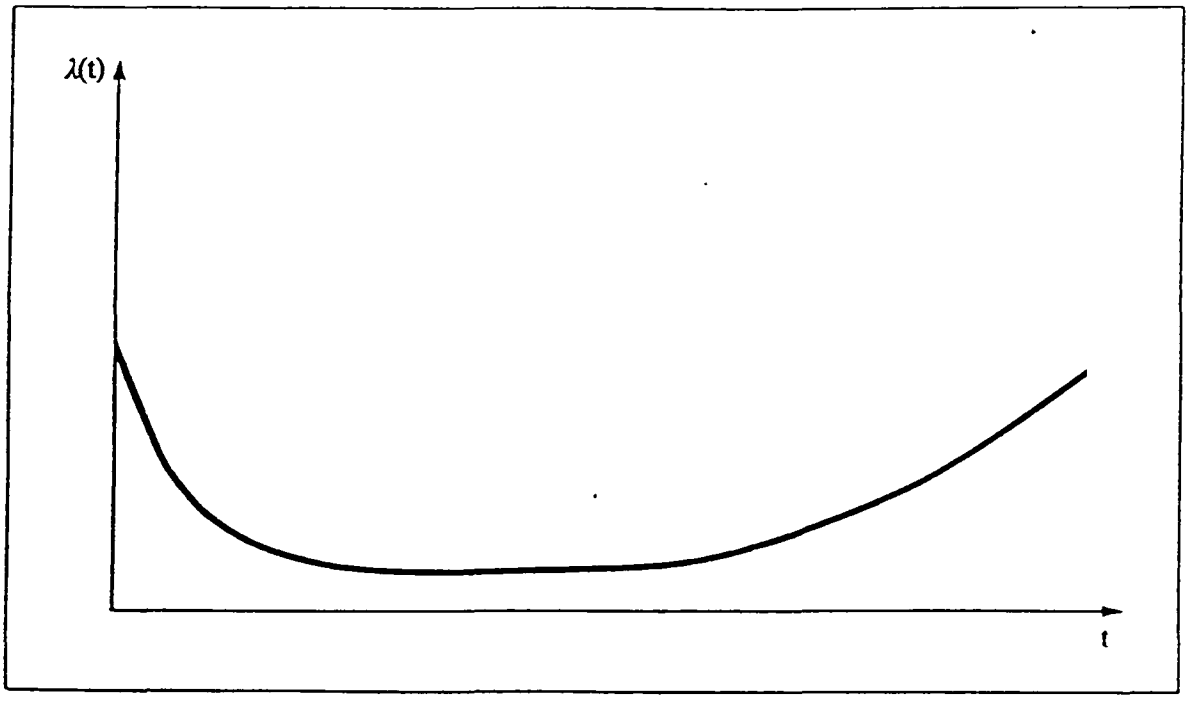

Figure 3.2.5: Hazard function of human mortality, where t=time. (Source: Blossfeld, Hamerle and Mayer, 1989)

The relationship between the survivor function and the hazard function (the inverse of the above situation) is obtained by integration of the hazard function which yields

$$S(x) = \exp\left(-\int_0^x \lambda(u)\, du\right) . \tag{3.2.9}$$

From equations (3.2.7) and (3.2.9), the probability density function is obtained as a function of the hazard rate

$$f(x) = \lambda(x) \cdot S(x) = \lambda(x) \cdot \exp\left(-\int_0^x \lambda(u)\, du\right). \tag{3.2.10}$$

The equations shown in (3.2.7) to (3.2.10) illustrate that each of the three measurements $f(x)$, $S(x)$, and $\lambda(x)$ may be used to describe the spatial durations. Thus, if one of these functions is known, it is always possible to derive both the other functions (Kiefer, 1988).

In most cases, event history analysis models the hazard rate, $\lambda(x)$, rather than $f(x)$ or $S(x)$. One reason to model the hazard rate is that, in substantive terms, it is important to consider the probability of a duration terminating, given that the duration has not ended at that distance. In addition, it is possible to model the effects of the present values of the explanatory variables on the hazard rate, even if this probability is dependent on certain spatially-varying explanatory variables. Without modelling $\lambda(x)$, it is difficult to make reasonable assumptions about the way $f(x)$ or $S(x)$ depend on previous values of time-varying explanatory variables. In general, the hazard rate illustrates the notion that individuals, or adopter location events, may differ in their

hazard rates, where events associated with large hazard rates are more likely to experience an event at shorter nearest neighbour durations, whilst those with small hazard rates are likely to experience the event at longer nearest neighbour durations (Yamaguchi, 1992).

Two major analytical methods are used for analyzing hazard rates: fully parametric models and partially parametric models. Recall that fully parametric approaches involve the use of a parametric distributional form to describe the nature of duration dependency. Partially parametric approaches are generalizations of parametric models, originally proposed by Cox (1972). They are referred to as partially parametric because the function describing the duration dependency is not specified *a priori*, whilst a functional form is specified for a regressive component in the model. Both fully and partially parametric models estimate the effects of explanatory variables on hazard rates. In contrast, nonparametric methods do not specify a relation between the hazard rate and explanatory variables, and are not used in this thesis (Yamaguchi, 1992). The subsequent discussion outlines some technical details of the fully parametric approach while the partially parametric approach is discussed in section 3.2.4.

Several parametric distributions could be considered for the duration or "distance to neighbouring event", X, distribution. Four widely used distributions in event history modelling are used in this thesis, the exponential, Weibull, Gompertz and log-logistic distributions. One of the most commonly applied distributions is the *exponential* distribution. In this case, the hazard rate is constant and represents the situation where no duration dependency exists. In mathematical terms, the constant hazard is represented

by

$$\lambda(x) = \lambda, \textit{ where } x \geq 0,\ \lambda > 0. \tag{3.2.11}$$

The respective survivor and density functions are

$$S(x) = \exp(-\lambda x), \tag{3.2.12}$$

$$f(x) = \lambda \exp(-\lambda x), \tag{3.2.13}$$

and the graphical representation of these formulations is shown in Figure 3.2.6.

A constant hazard in time may be an unrealistic assumption in many cases such as employment history where time invested in a job is likely to cause a decreasing hazard rate if the event is employer change. Similarly, with spatial durations, the constant hazard rate is equivalent to a completely spatial random distribution of events where the location of one event has no effect on the location of its neighbour - a very rigid assumption. Thus, a closely related distribution known as the *Weibull* model is often used to analyse durations. The hazard now has the form

$$\lambda(x) = \lambda\alpha\ (\lambda x)^{\alpha-1} \qquad (x>0) \tag{3.2.14}$$

where λ and α are parameters greater than zero (Kalbfleisch and Prentice, 1980). The respective survivor and density functions are

$$S(x) = \exp(-(\lambda x)^{\alpha}) \tag{3.2.15}$$

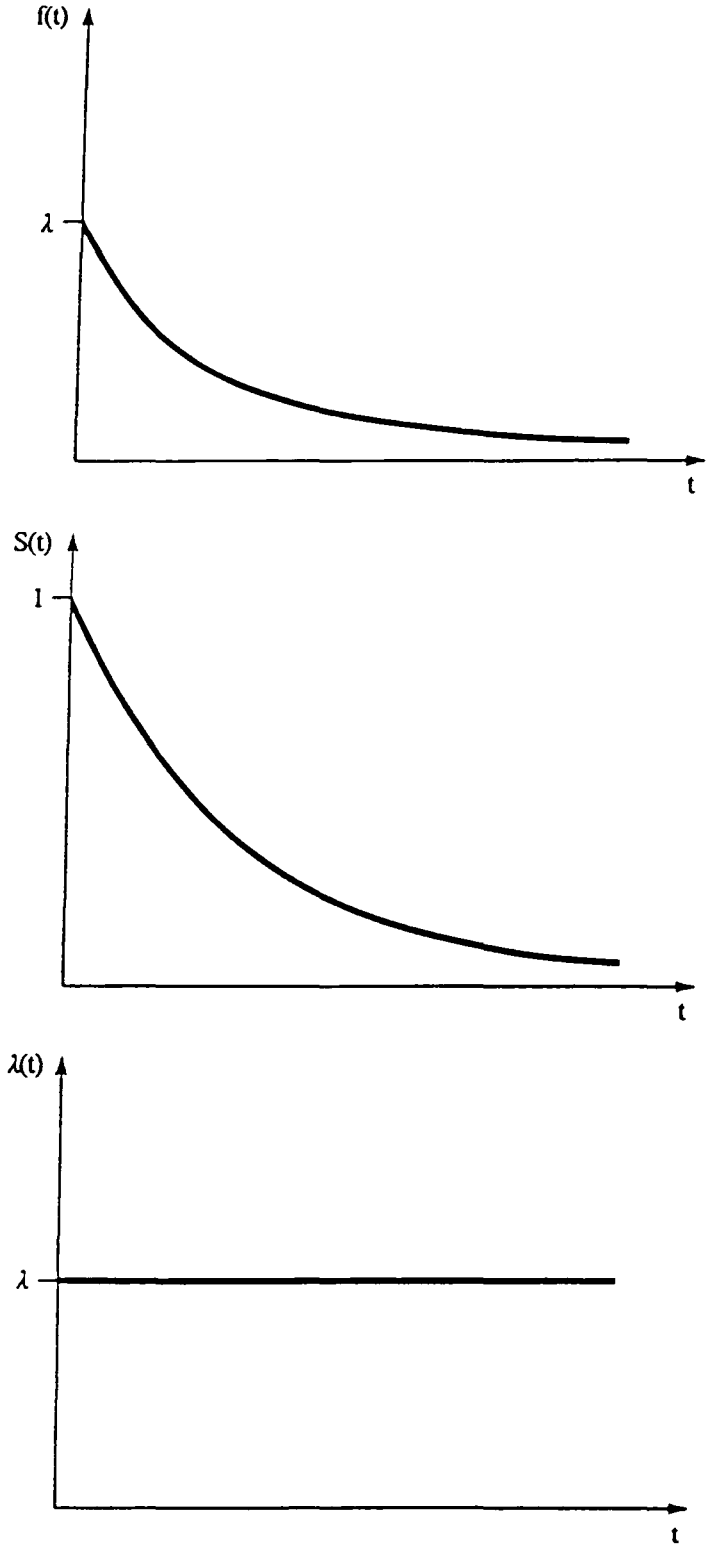

Figure 3.2.6: Density function, survivor function and hazard rate of the exponential distribution, where t=x. (Source: Blossfeld, Hamerle and Mayer, 1989)

$$f(x) = \lambda\alpha(\lambda x)^{\alpha-1}\exp(-(\lambda x)^{\alpha}). \qquad (3.2.16)$$

The hazard function of the Weibull distribution increases monotonically if $\alpha > 1$, decreases monotonically if $\alpha < 1$, and reduces to the exponential distribution when $\alpha = 1$ (i.e., constant hazard rate). The Weibull model is quite flexible and therefore adaptable in many situations (Figure 3.2.7). However, it does not allow for a *U*-shaped or inverted *U*-shaped hazard rate. This implies that the hazard may either decrease or increase with duration, but may not change direction (e.g., cannot increase and then decrease) with distance (Allison, 1984).

Another commonly used alternative to the exponential model is the *Gompertz* distribution. In this case, a Gompertz model for the distance to a neighbouring event, *X*, distribution yields the following hazard rate

$$\lambda(x) = \lambda_0 e^{\delta x}. \qquad (3.2.17)$$

The survival and density functions of the Gompertz model are

$$S(x) = \exp[(-\lambda_0/\delta)(e^{\delta x}-1)] \qquad (3.2.18)$$

$$f(x) = (\lambda_0 e^{\delta x})\exp[(-\lambda_0/\delta)(e^{\delta x}-1)] \qquad (3.2.19)$$

where δ is a positive or negative constant. Like the Weibull model, the Gompertz model is a monotonic distribution that allows the hazard to increase or decrease with duration (Figure 3.2.8).

In contrast to the constant and monotonic increasing or decreasing distributions discussed so far, the *log-logistic* is a non-monotonic distribution which accommodates a

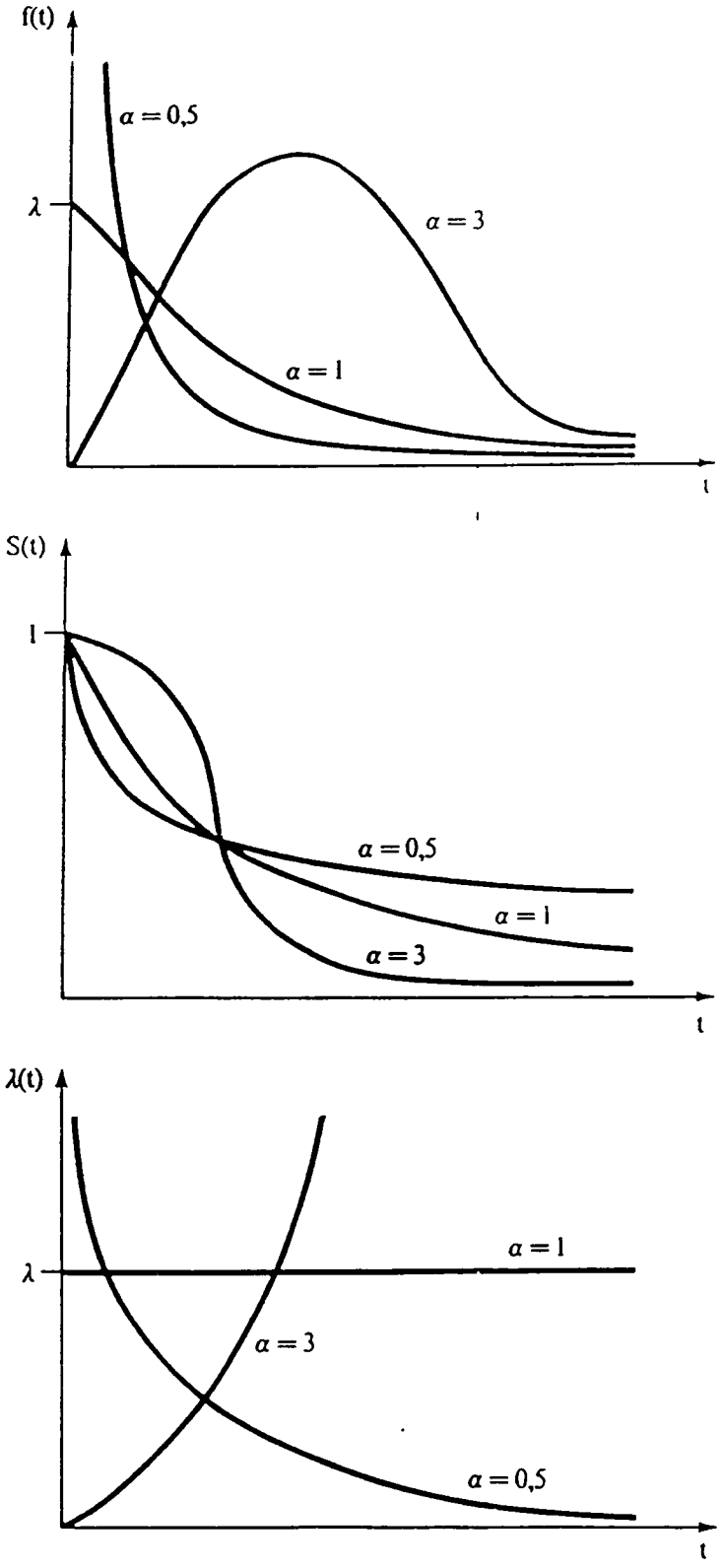

Figure 3.2.7: Density function, survivor function and hazard rate of the Weibull distribution, where t=x. (Source: Blossfeld, Hamerle and Mayer, 1989)

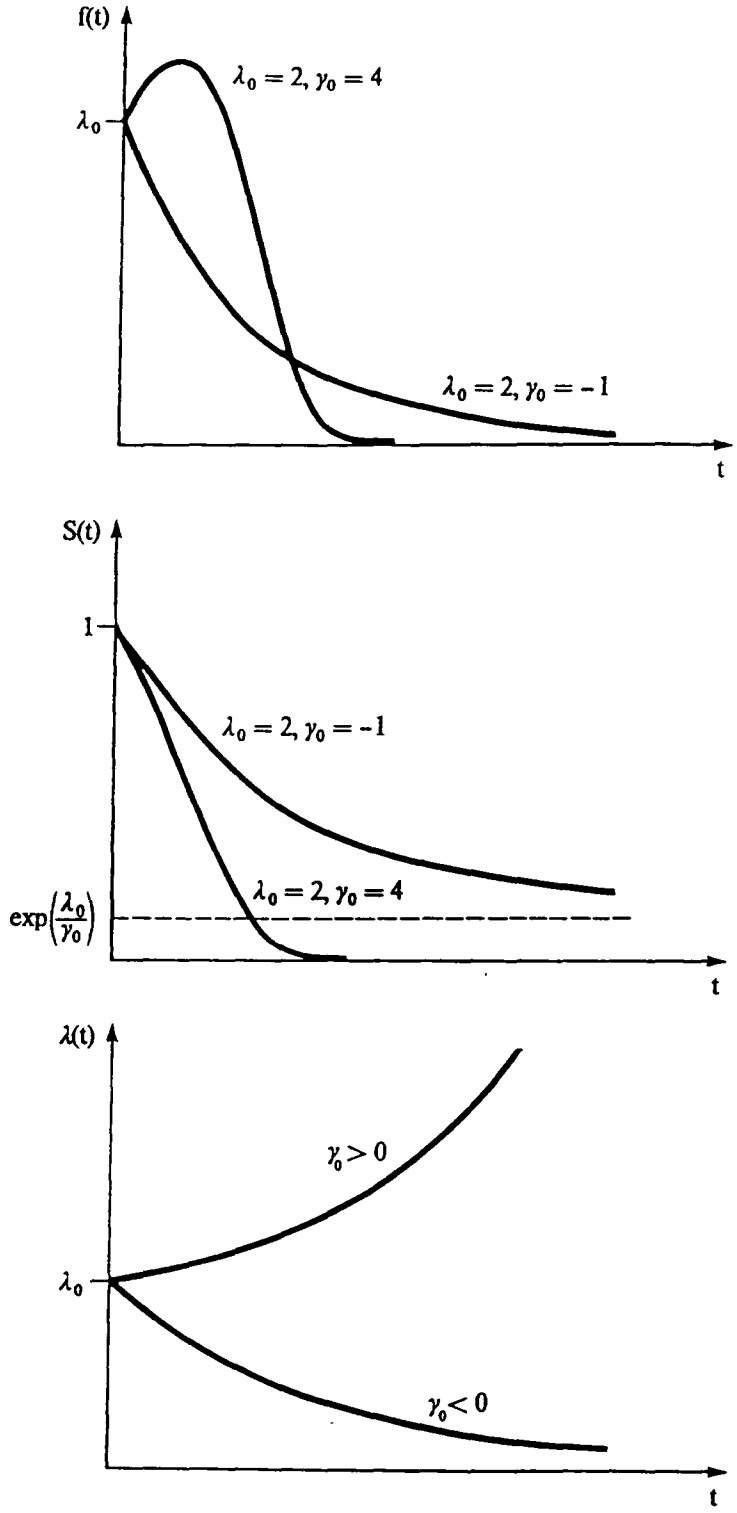

Figure 3.2.8: Density function, survivor function and hazard rate of the Gompertz distribution, t=x and $\gamma_0=\delta$. (Source: Blossfeld, Hamerle and Mayer, 1989)

hazard rate that changes direction (e.g., may increase and then decrease) with duration (Figure 3.2.9) (Kiefer, 1988). The log-logistic is specified in terms of log X with two parameters, λ and σ, greater than zero. In mathematical terms,

$$\ln X = \mu + \sigma \omega \tag{3.2.20}$$

and a logistic distribution for ω is specified with the density function

$$f(\omega) = \frac{\exp(\omega)}{[1+\exp(\omega)]^2} \quad . \tag{3.2.21}$$

By setting $\lambda = e^{-\mu}$ and $\alpha = \sigma^{-1}$, the density function for the log-logistic distribution is obtained as

$$f(x) = \lambda \alpha (\lambda x)^{\alpha-1} [1+(\lambda x)]^{-2} \quad . \tag{3.2.22}$$

The hazard and survivor functions are given by

$$\lambda(x) = \frac{\lambda \alpha (\lambda x)^{\alpha-1}}{1+(\lambda x)^{\alpha}} \tag{3.2.23}$$

$$S(x) = \frac{1}{1+(\lambda x)^{\alpha}} \quad . \tag{3.2.24}$$

For $\alpha > 1$ the hazard first increases with duration, then decreases. If $0 < \alpha \leq 1$, the hazard function decreases with duration.

The choice of a particular model involves deciding between theoretical appropriateness, empirical evidence and mathematical convenience. Given the unique nature of analyzing spatial point patterns using event history models and the availability of custom estimation routines in the software package used in this research, the approach

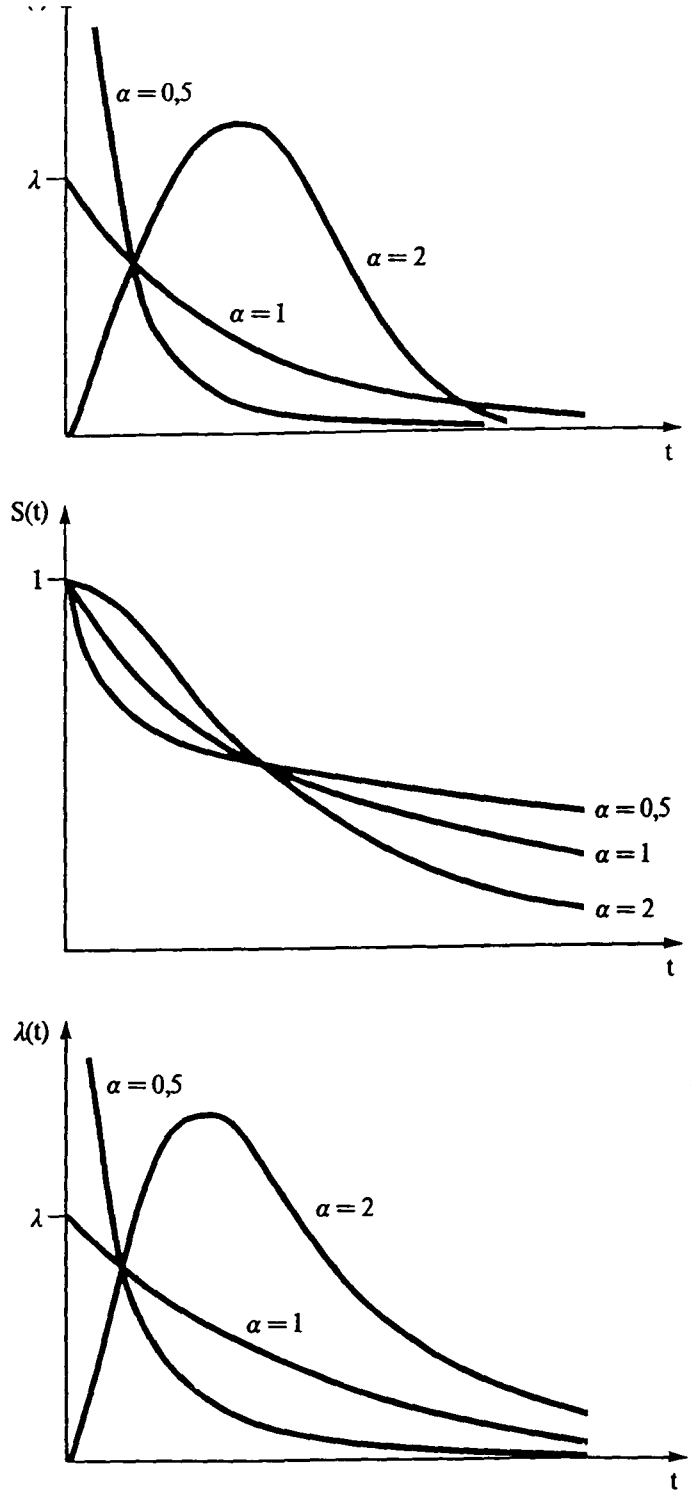

Figure 3.2.9: Density function, survivor function and hazard rate of the log-logistic distribution, where t=x. (Source: Blossfeld, Hamerle and Mayer, 1989)

taken in this thesis is to estimate, on an individual basis, the various parametric models outlined above, as well as the partially parametric Cox model . In this way, comparisons can be made as to how well commonly used distributional forms fit the spatial durations of the spatial diffusion point pattern data used in this analysis. Alternatively, substantive theory or well-known graphical methods such as an empirical graph of the survivor function using the "product-limit" (or *Kaplan-Meier* estimate) can be used to gather information on an appropriate model. Plots of estimated survivor functions provide useful depictions of event history data, as well as information on the underlying distribution (Blossfeld, Hamerle and Mayer, 1989; Lawless, 1982).

3.2.4 Regressive Event History Approaches

As was noted in section 3.2.1, one of the most important developments in event history analysis has been the incorporation of explanatory variables. The introduction of a regressive component to event history analysis is a reflection of the desire to use maximum information on units of study (e.g., event locations), including explanatory variables normally collected within an event history data set, to ascertain the quantitative influence of these exogenous or endogenous variables on the hazard rate. Here, the sign of the regression coefficient(s) indicates the directional effect explanatory variables have on the likelihood of a duration ending. Regressive event history models can be divided into two broad classes, namely *accelerated failure-time* models and *proportional hazards* models, depending on how the explanatory variables are presumed to affect the baseline hazard rate (i.e., the hazard without explanatory variables). In the case of accelerated

failure-time models, the effect of the explanatory variables is to alter the distance at which a duration comes to an end. In other words, a baseline hazard rate is assumed, $\lambda_0(x)$, and the regression variables increase or decrease the distance to event occurrence, X. In contrast, proportional hazards models represent the class of regression models in which the effect of explanatory variables is to multiply the hazard function itself by a scale factor, therefore, shifting the hazard rate up or down. In this thesis, both classes of models are estimated.

First, we consider the class of proportional hazards models. These models include the Cox (1972) model which differs from the fully parametric approaches defined earlier because no baseline hazard function is specified, yet estimates of the effects of explanatory variables are readily available. In the general form of the proportional hazards model, explanatory variables enter the model as a vector denoted z while β represents the corresponding parameter vector. The hazard rate of the model is then

$$\lambda(x|z) = \lambda_o(x)\exp(z'\beta) \; . \tag{3.2.25}$$

The baseline hazard rate, $\lambda_0(x)$, is usually specified by a specific parametric form, and in this thesis the Weibull and Gompertz distributions are used. In contrast, the Cox proportional hazards model does not specify the baseline hazard rate which allows greater flexibility for modelling, but requires *partial likelihood* estimation (as opposed to maximum likelihood used with the fully parametric models, see below). This so-called *partially parametric* approach suits an analysis where the main concern is with the effects of explanatory variables, while the dependence on distance is of little interest. For example, in biomedical applications using temporal durations, the interest focuses on the

effects of alternate treatments (the explanatory variables) rather than the dependence of the hazard on time.

The proportionality of the hazard rates means that the quotient

$$\frac{\lambda(x|z_1)}{\lambda(x|z_2)} = \exp((z_1 - z_2)'\beta) \tag{3.2.26}$$

for two event locations with explanatory variables z_1 and z_2 does not depend on duration (x). That is, for any two event locations, the ratio of their hazards is a constant over distance, x. Allison (1984) notes that this is not a crucial feature for either fully parametric models or the Cox model because the hazards cease to be proportional as soon as one introduces spatially-varying explanatory variables. He also adds that even when the proportionality assumption is violated, it is, nevertheless, often a satisfactory approximation, and analysts should be more concerned with any omitted explanatory variables then nonproportionality.

As noted above, by assuming a certain parametric form of the baseline hazard rate, the fully parametric proportional hazards models are obtained. In all the models used in this thesis, the non-negative functional form, $\exp(z'\beta)$, is used to specify the effects of the vector of covariates, once again z is a vector of explanatory variables and β is the corresponding parameter vector. For example, the Weibull is specified as

$$\lambda(x|z) = (\exp(z'\beta))^{\alpha-1}\alpha\exp(z'\beta) \tag{3.2.27}$$

and the exponential model is obtained from equation (3.2.27) by setting $\alpha=1$. Fully parametric proportional hazards models are used to examine the effects of both duration and explanatory variables on the hazard rate by comparing results obtained by partially parametric formulations, where no assumption on the form of the baseline hazard is made, *a priori*.

The class of accelerated failure time models, also known as "log-linear models," are obtained from the general form in equation (3.2.20) for the random variable representing the duration, X. Recall that the error variable, ω, can be specified as the logistic distribution which produces the log-logistic model. Adding the regressive component as before, the hazard rate of the log-logistic model is

$$\lambda(x \mid z) = \frac{\exp(z'\beta)\ \alpha\ (\exp(z'\beta)x)^{\alpha-1}}{1 + (\exp(z'\beta)x)^{\alpha}} \qquad (3.2.28)$$

and allows the hazard rate to be nonmonotonic, unlike the exponential, Weibull and Gompertz models. Given the unexplored nature of duration analysis in the spatial domain, a model with a nonmonotonic hazard rate may be appropriate for the spatial durations associated with complex spatial processes. Here, it is also prudent to note the multiplicative effect of the explanatory variables on x (or additively on lnX) in the accelerated failure time model, rather than on the hazard rate as in the proportional hazards formulation.

The fully parametric models specified above, both the proportional hazards and accelerated failure time versions, are estimated using *maximum likelihood* procedures. Maximum likelihood estimation (MLE) has become a standard procedure for parametric

model estimation and detailed explanations can be found in Blossfeld, Hamerle and Mayer (1989) or Kalbfeisch and Prentice (1980). Since MLE is used to estimate the parametric models in this thesis[1], a brief discussion of some of its properties is in order. MLE requires a parametric specification of the baseline hazard function, $h_0(x)$ and can accommodate both censored and uncensored observations (Tuma and Hannan, 1984; Yamaguchi, 1992). Also, ML estimators have been shown to possess superior asymptotic properties in large samples under fairly general conditions for the probability distribution function of the random variable, in this case distance, X (Tuma and Hannan, 1984). Thus, as the sample size tends to infinity, the MLEs are unbiased, normally distributed and efficient (i.e., contain minimal variance). Finally, MLE is not a computationally strenuous method as opposed to alternative methods of estimation such as Ordinary Least Squares for event history analysis (Allison, 1984).

The partially parametric Cox model, however, is estimated by partial likelihood. Recall that the partially parametric approach is useful when there is little theoretical guidance on the choice of an appropriate parametric baseline hazard, but unbiased and consistent estimates on the explanatory variables coefficients are required. The term partial likelihood refers to the fact that estimates based on this function depend only on the ordering of the lengths of durations and do not use information on the exact distance to a neighbouring event. This loss of information in partial likelihood estimation may be important if the major concern is with dependence of the hazard rate on distance, but

---

[1]MLE is available in most event history software packages, such as **RATE** (Tuma, 1979), **SURVREG** (Preston and Clarkson, 1983) and the package used in this thesis, **LIMDEP** (Greene, 1990). Chapter Four provides an overview of all the software used in this research.

where the concern is with the effects of the explanatory variables, partial likelihood estimation has been shown to be very efficient (Lawless, 1982; Tuma, 1982; Tuma and Hannan, 1984).

The partial likelihood approach proposed by Cox (1972, 1975) can be used for estimating the coefficients of β in the proportional hazards model [as in equation (3.2.25)], in the presence of censoring, without specifying the form of the baseline hazard function λ_0. We let a set of k durations be ordered from the shortest to the longest, where x_i is either the censored distance or the distance to a nearest neighbour, $x_1 < x_2 < \ldots x_k$ and let R_j denote the set of events with associated durations smaller than x_j. An indicator, in this thesis denoted CEN, is set to 1 if the duration ends at a nearest neighbour, or 0 if the distance ends at the boundary (i.e., the distance is censored). In partial likelihood estimation, the likelihood function is the product of likelihoods for all completed (or uncensored) durations only, not the likelihoods for all the observed durations. This method is best illustrated with an example derived from Allison (1984).

Given $k=10$ (i.e., 10 durations), for $n=10$ event locations ranked from shortest to longest, and the 4th, and 7th ranking durations are censored (i.e., measured to the boundary), this means $n=8$ events have uncensored durations. If the 1st ranking duration (i.e., the shortest) ended at distance x_j, then all the other events in this example were at risk of their duration ending (meeting a nearest neighbour or the boundary) at this distance from their location because it is the shortest. Thus, the likelihood that the duration ended for event $n=1$ rather than to one of the other 9 events is derived by taking the hazard rate of event $n=1$ at distance x_j and dividing by the sum of the hazards

for all the events at risk at distance x_j. However, the likelihood function for event $n=4$, which is censored, is not constructed, and this is also the case for event $n=7$. Basically, the censored durations are incorporated in the denominator of the likelihood as part of the events "at risk" of a duration ending, but since they are not complete observations, they do not have a corresponding likelihood function.

Formally, the conditional probability that a particular duration (indexed by i) is the duration which ends at x_j is

$$P_j(i|\boldsymbol{\beta}) = \frac{\exp[\boldsymbol{\beta} z_i(x_i)]}{\sum_{k \in R(z_j)} \exp[\boldsymbol{\beta} z_k(x_j)]} \tag{3.2.29}$$

where $z_k(x_j)$ is the regressive component of the kth duration, evaluated at x_j. Estimates of $\boldsymbol{\beta}$ are based on the likelihood function that is obtained as the product of these conditional probabilities,

$$L(\boldsymbol{\beta}) = \prod_{i \in I} \left[\frac{\exp[\boldsymbol{\beta} z_i(x_i)]}{\sum_{k \in R(x_j)} \exp[\boldsymbol{\beta} z_k(x_j)]} \right] \tag{3.2.30}$$

where $\boldsymbol{I}$ is the set of completed (i.e., uncensored) durations. The maximum of this likelihood function provides estimates that are asymptotically normally distributed under very general conditions (Kalbfleisch and Prentice, 1980). The basic notion here is that, in the absence of all information about the baseline hazard, only the order of the durations provides information about the unknown coefficients (Kiefer, 1988).

The inclusion of explanatory variables into event history models is important for two reasons. First, it is consistent with explanatory data analysis where the estimation

of the coefficients on the structural variables is intrinsically valuable. Second, the inclusion of explanatory variables is important in correctly identifying distance (or duration) dependency in the hazard rate. This issue concerns the distinction between true duration dependence and spurious duration dependence arising from variation in hazard rates across event locations (i.e., environmental heterogeneity). In general, if the substantive concern is with the dependence of the hazard rate on distance, then fully parametric regression methods may be more appropriate then the partially parametric approaches which do not provide direct assessment of the baseline hazard rate. The main attraction of the partially parametric Cox model is that it allows the effects of explanatory variables to be examined under general conditions for the baseline hazard rate thereby reducing the risk of model misspecification from incorrect parametric assumptions. This feature has lead to widespread application of this method in the analysis of time intervals and, recently, to the analysis of distance measurements by Odland and Ellis (1992) that is discussed below.

3.2.5 Settlement Pattern Analysis with Proportional Hazards Models

In their application of proportional hazards models to the analysis of a spatial point pattern (i.e., settlements), Odland and Ellis (1992) note that spatial point process models maintain some restrictive assumption on either the interdependence or heterogeneity in the density of events which leads to ambiguous conclusions about the heterogeneity of the study area. For example, under CSR, the assignment of events to locations is homogeneous and independent, whilst alternative models incorporate some

form of interdependence among the locations of events such as contagion or inhibition. Models allowing for heterogeneity in the density of events are also available, but maintain restrictive conditions on event interdependence (Diggle, 1983). Typically, these spatial process models imply particular frequency distributions for the distances separating points and comparisons of observed inter-event distances with the frequencies implied by a model would reveal the operation of a particular process. However, such comparisons provide only ambiguous evidence about alternative processes, because the frequency distributions implied by these models depend on specific conditions for both heterogeneity and interdependence and deviation from the implied frequency distributions may result from the failure of either condition.

Odland and Ellis (1992) turn to the partially parametric proportional hazards model as an alternative to spatial process models. In this respect, they are able to measure trends in the nearest-neighbour durations between events in space, rather than the interdependence among event locations, and, therefore, a null hypothesis of homogeneity could be tested without an explicit specification for the interdependence of event locations. Recall that the proportional hazards model can be specified in both the fully parametric and partially parametric format. If the main concern is with testing the location of an event in space based on the effects of explanatory variables, a model could be specified without assuming a particular form for the spatial dependence. Thus, the proportional hazards model allows for the unbiased estimation of spatial heterogeneity effects, unlike the standard spatial process models (Odland and Ellis, 1992).

The straightforward application to settlement spacing by Odland and Ellis (1992)

included testing for CSR by comparing the partially parametric proportional hazards model with the constant hazard Weibull model shown in equation (3.2.27), which is equivalent to the exponential distribution. In this regard, a process in which events are located randomly and independently within a two-dimensional plane leads to an exponential distribution for the distances between neighbouring events. Therefore, evidence for or against a random and independent assignment of nearest neighbour distances may be sought by comparing a set of observed distances to those implied by the exponential distribution. Spatially varying explanatory variables were incorporated by including both the distance north-south and east-west from the south-east corner of the study area, Nebraska. Results indicated that a definite east-west trend existed in the settlement pattern, with the spacing of settlements increasing toward the western part of the state. This finding confirmed the authors hypothesis based on the obvious east-west trend available by visual inspection of the spatial point pattern.

Further analysis considered any interdependencies between these settlement locations by comparing survivor functions of the proportional hazards model with the survivor functions of a model of Complete Spatial Randomness (CSR). Recall that CSR incorporates the assumptions of homogeneity and independence. Comparisons of these survivor functions at a series of areas defined by isolines oriented orthogonally to the direction of change (in this case, north-south) indicate a tendency towards uniform spacing at these small spatial scales (Figure 3.2.10). In other words, there is less variation in the distance intervals between nearest neighbours (i.e., fewer very large or very small measures) than would be expected under CSR implying a competition for

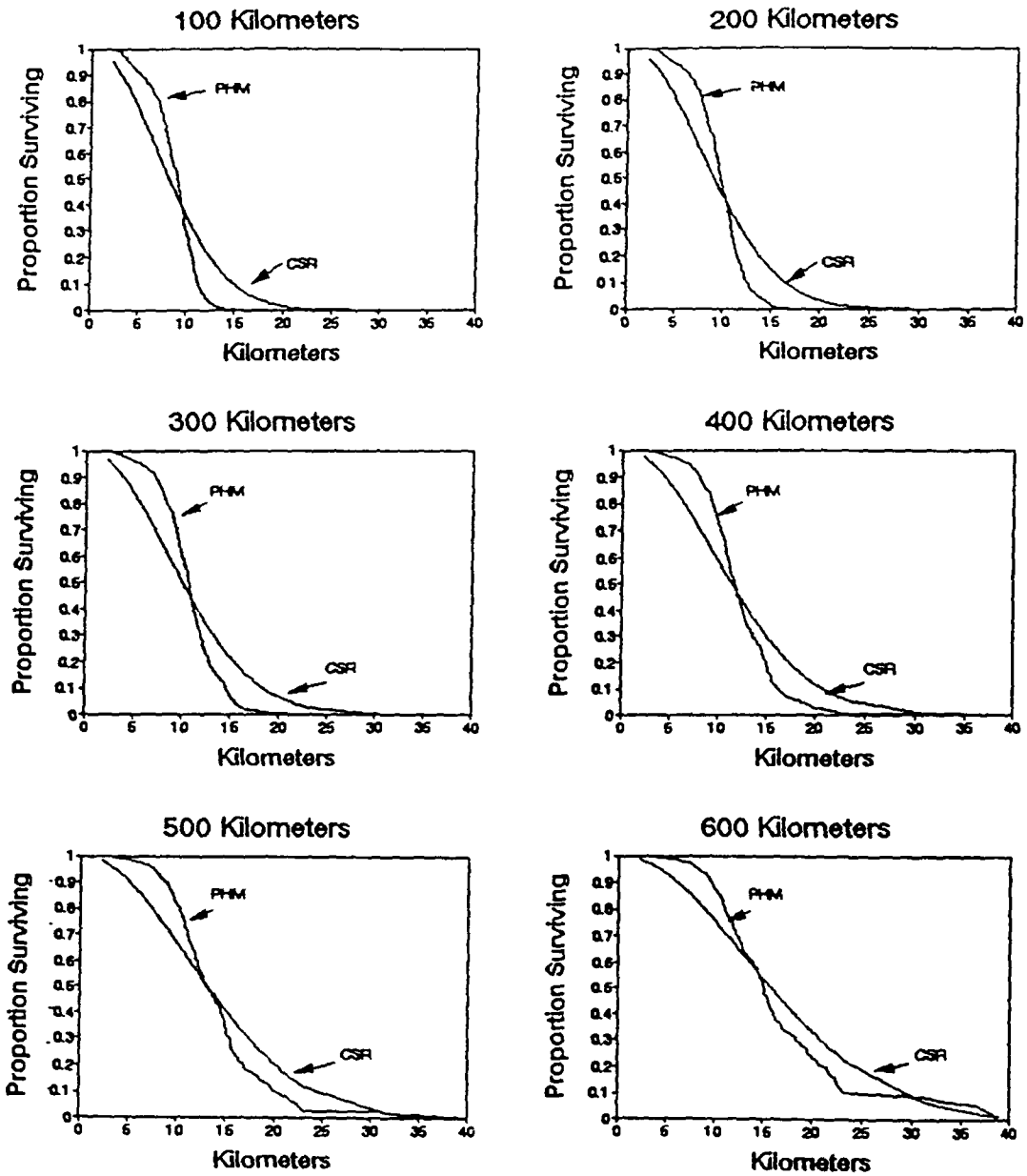

Figure 3.2.10: Comparisons of survivor plots for the Proportional Hazards Model (PHM) and the model of Complete Spatial Randomness (CSR); 100 to 600 kilometers west, Nebraska. (Source: Odland and Ellis, 1992)

market areas and uniform spacing (Odland and Ellis, 1992).

3.3 Conclusions

In this chapter, basic definitions and statistical concepts of event history analysis were introduced. More importantly, however, this chapter sought to translate concepts associated with the analysis of time intervals into useful methods for the analysis of spatial duration (or distance) measurements. Sources of spatial variation such as environmental heterogeneity of the process environment and the individual event, along with duration dependency were defined. The emphasis in this chapter was clearly on regressive models and the measurement of censored durations. The need to evaluate the effects of explanatory variables on a duration measured from an event location to its nearest neighbour allows some assessment of the spatial diffusion process of interest in this thesis. Chapter Two identified general explanatory variables, such as the sales agent and homophilous individuals, that may influence the spread of information about an agricultural innovation to other farmers and, presumably, lower their resistance to adoption. The next chapter defines the explanatory variables used in this thesis and provides hypotheses concerning their possible effects.

Besides the obvious utility of regressive approaches, the notion of spatial censoring has intuitive appeal for the analysis of a series of events in space. In Chapter Two, the problems associated with boundary or edge effects, notably the bias toward longer mean nearest neighbour distances, was discussed. In this chapter, "right-censoring" was redefined in spatial terms so that edge effects could be incorporated into

model parameter estimation. In this sense, spatial censoring occurs when an event in space is closer to the boundary then to a nearest neighbour (i.e., a boundary event) is associated with a duration measurement defined as the shortest distance from itself to the boundary. This censored duration is incorporated into the model estimation procedure and provides certain knowledge that no event is closer to that event with the censored duration than the boundary is. To properly assess the usefulness of measuring censored durations, the results should be compared to results from conventional nearest neighbour durations. Therefore, spatial durations that include censored durations and those that are strictly uncensored (where the boundary event must now find a neighbour within the study area) are analysed separately.

Chapter Three also summarized the analysis by Odland and Ellis (1992) based on proportional hazards models. From the review of relevant methods provided in this chapter, several other areas of investigation are clear with respect to the usefulness of event history methods for the analysis of spatial point patterns. Two have already been mentioned above. First, this thesis attempts to measure the effects of explanatory variables on nearest neighbour durations associated with a diffusion process. Second, nearest neighbour durations that may terminate outside the boundary because they are closer to the boundary then to a neighbouring event are incorporated as spatially censored durations. These censored durations are analyzed separately from uncensored or conventional durations whose nearest neighbours must all lie in the study area. Also, a wider range of distributional forms for the baseline hazard rate are compared to assess their fit with the nearest neighbour durations, as opposed to Odland and Ellis' (1992) use

of only the exponential model. Spatially-varying explanatory variables are not included in this thesis, although Odland and Ellis' (1992) method of incorporating these variables is possible. However, the spatial point pattern in this thesis has no obvious trends in the spacing of events that can be associated with any explanatory variable, as was the case in the spatial point pattern used by Odland and Ellis (1992). The next chapter presents the data and analysis outline.

THIS STUDY IS CONCERNED WITH THE ANALYSIS OF A SPECIFIC GEOGRAPHIC AREA; ITS OBJECT IS TO DEAL WITH THE DIFFUSION OF INNOVATIONS AS A SPATIAL PROCESS. THAT THE MATERIAL USED TO THROW LIGHT ON THE PROCESS RELATES TO A SINGLE AREA SHOULD BE REGARDED AS A REGRETTABLE NECESSITY RATHER THAN A METHODOLOGICAL SUBTLETY.

HÄGERSTRAND

CHAPTER FOUR

DATA AND RESEARCH METHODOLOGY

4.1 Introduction

This chapter seeks to present the data and methodology under investigation in this thesis. The data consists of a spatial point pattern derived from the diffusion of an agricultural innovation (the **A.O. Smith Harvestore Feed Crop Storage System**) in a predominantly rural area between April, 1963 and September, 1986. In particular, the data includes the precise location of each adopter farm along with a series of explanatory variables associated with that farm and/or location. This data is analyzed in terms of the spatial durations between adopter locations (i.e., *events*) and departs from other spatial point pattern and spatial diffusion studies through the event history approach taken. In addition, although the methodology adopted in this research includes variables based on traditional diffusion of agricultural innovation studies, it does not concentrate on communication patterns cognate with studies done by rural sociologists.

This investigation attempts to further the research by Odland and Ellis (1992) in applying event history models to the analysis of spatial point patterns. In this respect, this thesis is methodological in nature with an emphasis on the utility of an event history

approach to the study of spatial patterns and processes. The present analysis is similar to that of Odland and Ellis (1992) in that it utilizes a spatial point pattern which contains some inherent spatial dependence between event locations. However, it extends the previous research by exploring a wider range of parametric models, spatially censored durations, and a more extensive set of explanatory variables. The investigation also employs simulated spatial point patterns and graphical methods to test the empirical spatial point pattern against one generated by complete spatial randomness (CSR). In addition, the spatial data handling capabilities of Geographic Information Systems (GIS) are used to facilitate the construction of the simulated events and measure the durations.

4.2 The Innovation: The Harvestore Feed Crop Storage System

The "Harvestore System," also known as the *Cadillac* of silos, is a unique feed crop storage system that has a number of advantages over ordinary farm silos. A serious problem with feed crop storage in ordinary silos is that up to one quarter of the feed crop is lost through oxidation. Atmospheric temperature changes cause gases inside silos to expand and contract. This action exerts pressure on the silo structure which cannot be compensated for without allowing air to enter and contact the feed crop, so causing oxidation (DeTemple, 1971; Scorgie, 1973).

The major advantage of the Harvestore System is that it can be sealed air-tight to reduce feed crop loss through oxidation. The Harvestore structure is constructed of glass-fused-to-steel plates that are impervious to air. Inside the structure, pressure absorbing gas bags vested to the outside compensate for changes in atmospheric

temperature and pressure. A rise in outside temperature causes gases inside the structure to expand and push air out of the breather bags. Conversely, a decline in outside temperature causes gases inside the silo to contract and the breather bags are filled with air. Thus, the system, by controlling in and out air flow, compensates for pressure changes inside the Harvestore structure without allowing air to contact the feed crop.

In addition to the obvious advantage a Harvestore System provides to a farmer by significantly reducing feed crop loss through oxidation, it also gives the farmer greater flexibility in cropping and harvesting, and allows the farmer to increase both the quantity and quality of animal feed. In this regard, feed crops can be harvested early when moisture and protein content are high and stored in the Harvestore structure without the worry or cost of drying the harvest (DeTemple, 1970, 1971). *Double-cropping* with a winter crop and an early spring harvest is a possibility that allows the farmer to get an extra crop per year from the same acreage. For example, beef farmers use the Harvestore structure for alfalfa haylage (high moisture hay), but this could be topped with high moisture corn in the fall season. Thus, the farmer could grow and store all the feed necessary for optimum production from the livestock without having the added expense of commercial feeds, vitamin supplements or other additives (Scorgie, 1973).

Related to the notion of double-cropping is the automatic unloading from the bottom of the Harvestore structure - another advantage of a Harvestore System. The automatic unloading of the stored feed from the bottom means that it is not necessary to unload the structure before refilling. Ordinary silos, on the other hand, load and unload from the top, thus they must be emptied before refilling. In other words, bottom

unloading enables farmers to "feed out" the feed placed in first, as well as being able to feed out while filling the structure with new feed.

Basically, the Harvestore System can be considered an innovation because of three features discussed above. In summary, the Harvestore System is an innovative feed crop storage system in that it does three things of which no other silo is capable:

(1) It resists corrosion from feed acids on the inside as well as from the outside with its glass-steel encasing.

(2) It provides maximum protection from oxygen to preserve feed nutrients with a "breather bag" system suspended from the top of the structure that permits air to transfer in an out of the structure (in the bags) and compensates for pressure differences inside and outside the air-tight structure.

(3) It automatically unloads from the bottom allowing double-cropping and reduces the heavy labour involved in unloading/loading the silo at harvest.

Basic to understanding the diffusion of the Harvestore System, and the reasons for its acceptance to any farmer, is knowledge of the context in which the Harvestore was developed. The history of the Harvestore is discussed in detail by Scorgie (1973, 1992), but some key aspects of the Harvestore's establishment are worth noting here. After the Second World War, the expansion of industry and the subsequent increase in the demand for agricultural products led to farmers increasing production and taking advantage of economies of scale. The competition within farming intensified such that the alternative to expansion was elimination from profitable farming. Adding to this climate of change in farming was the growing importance of corn as the primary feed

upon which Ontario livestock producers have come to rely. Along with the "corn revolution," the demand for feed storage silo towers swelled dramatically. Thus, farming operations with the storage of hay and grain in a barn shifted to silos and automatic feed-handling systems. Most silo towers were built of poured concrete, cement blocks or cement staves. Through the 1960s, the number and average size (in diameter) of silos increased reflecting the competitive nature of farming wherein farmers cropped more land, bought more livestock and generally increased the scale of their farming operations (Scorgie, 1973).

The corn revolution extended the limits of commercial farming in North America through the introduction of "hybrid corn." Hybrid corn, which give different yields and mature at different rates than regular corn, can be used for different types of feed. In fact, corn is considered the best field crop for producing energy in livestock in terms of total calories (Scorgie, 1973). Also, corn may be fed to animals as corn silage which uses all the plant minus the roots (i.e., chopped leaves, stock, cob, and kernels are combined as a feed). Alternatively, corn may also be fed to livestock as grain by removing the protein rich kernels from the cobs. The kernels may be dried or stored as high moisture corn; they may be ground before using them as feed. The versatility of corn and its high energy factor is in sharp contrast to traditional winter feeds such as alfalfa, buckwheat, clover or legumes, which are cut, dried, cured and stored in the barn as hay or in compacted form as bales. However, hay does not have as many total digestible nutrients (TDN) as corn silage (Scorgie, 1992).

A major trend in southwestern Ontario farming since the 1960s has been the shift

to higher energy corn from the lower TDN feeds (Scorgie, 1973). The three main types of feed crop, namely corn, grain and hay, are used on hog, beef cattle and dairy cattle farms in southwestern Ontario. The townships in the study area contain farms that are livestock feed-growing and livestock raising although fruit and vegetable, tobacco, cash crops and marginal pasturing farm areas exist. This trend towards corn and "wet feed" in general, lead to the increased use of silos to store feed. Wet feed, including the high moisture content corn, simply refers to feed not put through a long, labour intensive drying procedure common with hay usage. Thus, alfalfa haylage (high moisture content hay) and high moisture corn revolutionized feed crop farming in general (Scorgie, 1973; 1992).

The commercial Harvestore silo first made its appearance in southern Wisconsin and northern Illinois in 1949. Developed originally by the A. O. Smith Corporation in 1944, the structure was constructed of a glass-fused-to-steel material invented by steel fabricators decades earlier. The resulting silo exhibits the surface characteristics of hard, smooth, corrosive-resistant glass and yet it has the structural strength of steel. Unlike the cement cast-in-place silos, the Harvestore silo could be dismantled and reassembled elsewhere because it was designed to be bolted together. It is possible to take apart a stave silo piece by piece and rebuild it, but the result may be inferior to the original structure in terms of strength and ability to keep oxygen from reaching the silage. Most unwanted silos, in fact, are removed from their position by dynamiting or through demolition by wrecking crane resulting in worthless cement or, at best, cement fill that must be removed. The Harvestore structure could be sold and rebuilt at another location

as opposed to the more permanent cement alternatives.

In addition to the unique features mentioned above, the Harvestore is a complete system. That is, a farmer deciding to buy an *oxygen free* or *sealed storage* Harvestore System received the silo structure, a foundation, a filler hood, pipes and an unloader. In contrast, companies associated with concrete or stave silos may build an extension on to an existing silo (even if it did not erect it originally), may add unloading and conveying equipment, or may construct other farm buildings. Hence, a farmer bought a complete unit which was ready to be used when feed was supplied rather than simply buying a "hollow tube." Along with the complete package, however, was a price tag two to three times more expensive than the same size concrete silo (Scorgie, 1973; 1992).

A final aspect of the Harvestore System to consider is the attention required of the farmer for proper use and care of the structure. In this context, the farmer must take several precautions including checking the moisture content in the feed being placed in the structure; closing the hatch after each filling to prevent oxygen from entering the tank which would otherwise defeat the purpose of oxygen free storage; inspecting the unloading equipment to prevent unnecessary breakdowns; and replacing the plate at the end of the unloader to keep oxygen from reaching the feed at the bottom of the structure. Likewise, servicing damaged or malfunctioning Harvestore's was an important function of the two companies that marketed the Harvestore during the study period (see below). Indeed, farmers whose livestock depended on being fed daily would be less inclined to adopt the Harvestore if fast, dependable service was not available.

4.3 The Study Data Set

The data set analyzed in this thesis consists of coordinate locations of 528 adopter farms of the **Harvestore Feed Crop Storage System** between 1963 to 1986 (termed events). The Harvestore was marketed in the study area by two companies: Reliable Farm Specialties from 1962 to 1966 and Ontario Harvestore Systems from 1967 onwards. The data was collected by Mr. E. K. Scorgie and details of its collection are provided in Scorgie (1992). The Harvestore data is particularly well suited to this analysis as it contains records of each adoption both in terms of the exact location, date of purchase, and a set of explanatory variables. In this regard, the data set could accomodate both temporal and spatial regressive analyses, although in this thesis, only the spatial aspect is explored. The diffusion of agricultural innovations using the spread of Harvestores as data has previously been investigated by DeTemple (1970, 1971) and Scorgie (1973). However, in this research, the focus is on the spatial durations between adopter locations (events) and the resultant spatial pattern, rather than investigating the complex and dynamic sociological and economic diffusion mechanisms that characterize these earlier studies.

The pattern formed by the locations of adopters in southwestern Ontario is shown in Figure 4.3.1 using **TransCAD** GIS software. This set of 528 adopter locations consists of all farms that were initial adopters of the Harvestore system and excludes farmers purchasing subsequent Harvestore structures at the same location. The locations of these adopters may have been influenced by several factors discussed previously in section 2.3.2 and these include various sources of interpersonal communication. In the

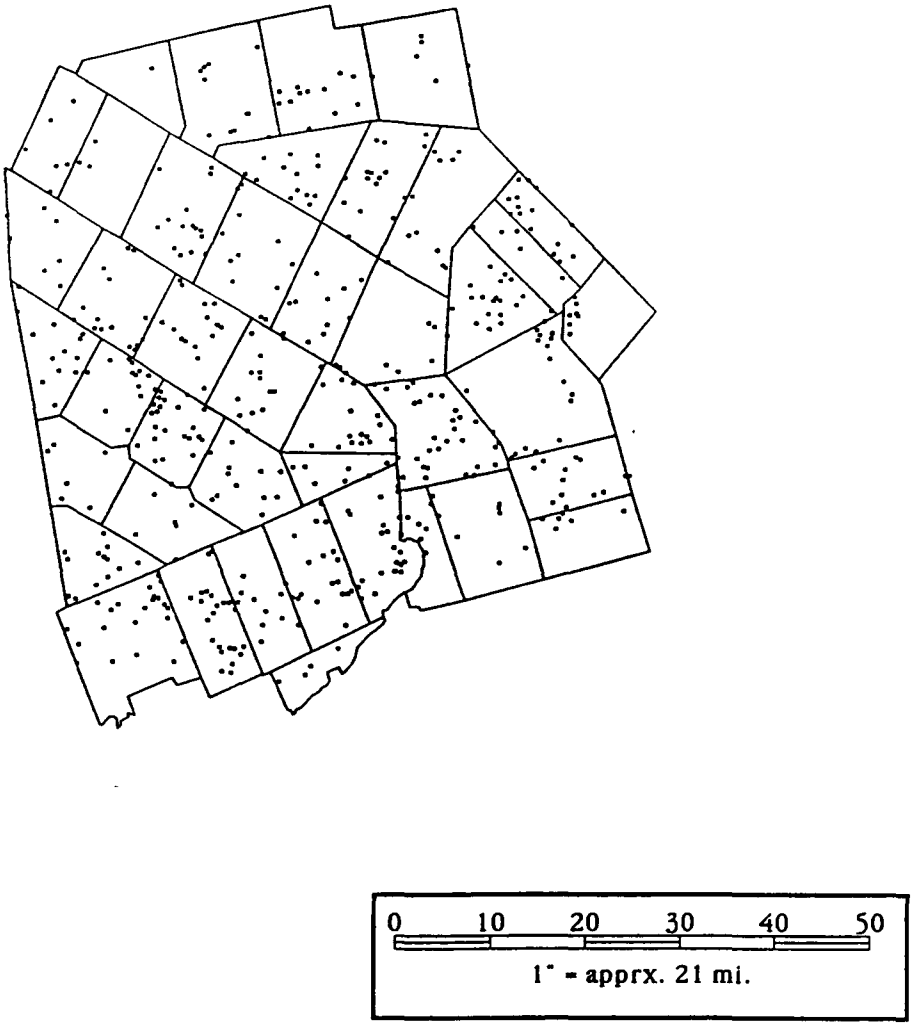

Figure 4.3.1: The spatial point pattern formed by Harvestore adopters.

following discussion, the dependent and independent variables are defined and some hypotheses concerning their influence on the parameter coefficients are presented.

4.3.1 Explanatory Variable Definitions

Recall that in this thesis, an *event* refers to the farm location of an initial purchaser (adopter) of the Harvestore system in the study area. This definition therefore excludes, as events, additional Harvestore purchases at this same location or relocated silos. A *duration*, to be modelled as the dependent variable, is defined as the Euclidean distance measured between nearest neighbour events. In other words, each event is associated with a distance measurement to the nearest Harvestore adopter. In addition, several characteristics of each event are available that may have some influence on the distance separating an event from its neighbour, and these may be defined as explanatory variables. Also, in regards to the durations of interest, recall from Chapter Three that two data sets of observed durations will be analysed: uncensored and censored durations. In the *uncensored* data set, each event has an associated duration that ends at a neighbouring event, while in the *censored* data set, some events (termed boundary events) will have associated durations that terminate at the boundary, but both types of duration can be influenced by the explanatory variables.

From the discussion in Chapter Two on the spatial diffusion of agricultural innovations, it is apparent that the spatial mechanisms of information flow are an important part of the learning and adoption process. A pivotal source of interpersonal communication and information flow in any diffusion study is the *change agent* of the

innovation (Rogers, 1983). In this thesis, the sales agent (SAL) is the change agent of interest and is one of the series of explanatory variables measured for each event (i.e., adopter farm). Over the 23 year period for which data was collected, a total of 42 sales agents sold Harvestore systems, at one time or another, in the study area. The sales agents worked for different periods of time, some for many years and some for only a few weeks, and the amount of agents working in the study area at one time was not consistent during the 23 year period (Scorgie, 1992). In addition, the sales agents generally worked in distinct spatial markets or sales territories within the study area, with the size of these territories varying considerably amongst agents and through time (Scorgie, 1992).

The above description of the sales agents role in diffusing the Harvestore innovation suggests a complex web of temporal and spatial mechanisms and using this variable as a meaningful regressor appears difficult. Nevertheless, the sales agents are viewed by Scorgie (1992) as the key ingredient in the spread of Harvestore systems, and must be included in an investigation of the diffusion process of Harvestore's for several reasons. First, the *motivation* which initiates information transmission is central to understanding the influence of sales agents on the learning and adoption process (Brown, 1981). In general, potential adopters of the Harvestore may obtain information by seeking it or by merely receiving it without solicitation. In the case of Harvestore sales agents, their motivation was the latter where agents looked for ideal farmers to approach with the idea of sealed storage. For example, a farmer contemplating a change in his/her operations to decrease expenses on feed or protein supplements, labour,

equipment or by increasing productivity, was a prime target of Harvestore sales proposals. This potential adopter was differentiated amongst the population of farmers by the sales agents through information gained by personal contact with farmers in their sales territory or by examining relevant census figures for hogs, dairy cattle, and beef cattle to calculate an index of the potential market for Harvestores by township (Scorgie, 1973; 1992).

Further, the Harvestore innovation, in broad terms, sought to make the conventional farm capital intensive rather than labour intensive, as is typical of the modern agricultural sector (Scorgie, 1973). For this reason, sales agents approached, for the most part, working farmers, generally under the age of 40, who have had some success in their farming operation, showing that they possessed managerial ability and that they would probably use the Harvestore silo properly. This was essential since the Harvestore carried a significantly higher price than conventional silos and, very often, financing was required. This being the case, the Harvestore should be able to help farmers reduce their labour input in the farm operation, at the same time reducing the feed costs and providing better quality feed for the livestock. To operate in this manner, the farmer had to plan to farm for at least more ten years - the realistic time required to repay the initial outlay of capital for the Harvestore (Scorgie, 1973). Thus, the definition of an "ideal" potential adopter of the Harvestore, from the sales agents point of view, is extended to young, innovative farmers, looking for a change on their farm: either to reduce labour or to increase profit, or both, **and** who could properly manage the Harvestore and its associated financing.

In some cases, it was easy for sales agents to find farmers who would want a change in their farming operations. For instance, sales agents were attracted to farmers whose barns had been destroyed by fire and whose farm operation would change through necessity. In other cases, farmers lacked storage facilities for the feed which they had grown, and they were forced to watch that feed deteriorate because it was not stored properly (Scorgie, 1973). However, sales agents were also confronted by farmers very resistant to change. In these instances, farmers with highly successful operations did not feel a need to invest more money in their already effective farming system. The Harvestore is, in fact, a system which does change a farmer's operation, from the crops he/she plants to the his/her harvesting and storing procedures, then on to the handling of his/her feed, and finally to his/her marketing of the agricultural product (Scorgie, 1973). Once again then, we may amend the notion of an "ideal adopter" to include farmers who are conscious of the necessity of getting the greatest amount of feed from the land that they have at their disposal, and so willing to change an established farming operation.

In summary, the sales agents look for prospects who are dedicated to farming and expect to farm for the foreseeable future. Since farming is their livelihood, they are interested in improving their position, as is any concerned business person. Farmers may improve their economic position by decreasing expenses or increasing productivity, or both, and the Harvestore provides this possibility for farmers. Therefore, the sales agents are looking for farmers who are contemplating a change. Secondly, a farmer must have shown from his/her past experiences that he/she is a good manager, not only

in his/her utilization of time and physical resources, but also in his/her ability to work with money. Both of these characteristics of potential adopters are linked to another marketing strategy by Harvestore sales agents - *the neighbourhood effect* (Scorgie, 1973).

In this regard, a successful operation displaying the distinctive blue coloured Harvestore silo was the best advertising to neighbouring farm operators any sales agent could ask for. On the other hand, selling a Harvestore to a farmer with poor managerial skills and a marginal farming operation could easily have lead to reposession of the Harvestore by a bank, and adverse publicity for the Harvestore. In the same vein, sales agents routinely visited owners in their sales territory up to four times each year, illustrating a willingness to service the Harvestore. This high level of customer service, according to Scorgie (1992), was also a neighbourhood effect selling strategy. Here, the sales agents hoped that satisfied Harvestor owners would "sell" the idea of sealed storage to other farmers (neighbours or social groups), while the agents themselves attempted to communicate information on the Harvestore silo to neighbouring farmers. In his earlier study, Scorgie (1973) noted that the greatest number of potential adopters were brought to the attention of sales agents through satisfied Harvestore owners, consistent with the neighbourhood effect of diffusion.

Given the importance of the sales agent as the "contagion" facilitating the farmer's decision-making process and potential acceptance of the innovation, the sales agent data was considered more closely. Visualizing the data as a histogram revealed an interesting dichotomy in the sales agent data (Figure 4.3.2). Here, we see that three sales agents (coded 5, 6, and 14) accounted for approximately 62% of Harvestore adoptions and each

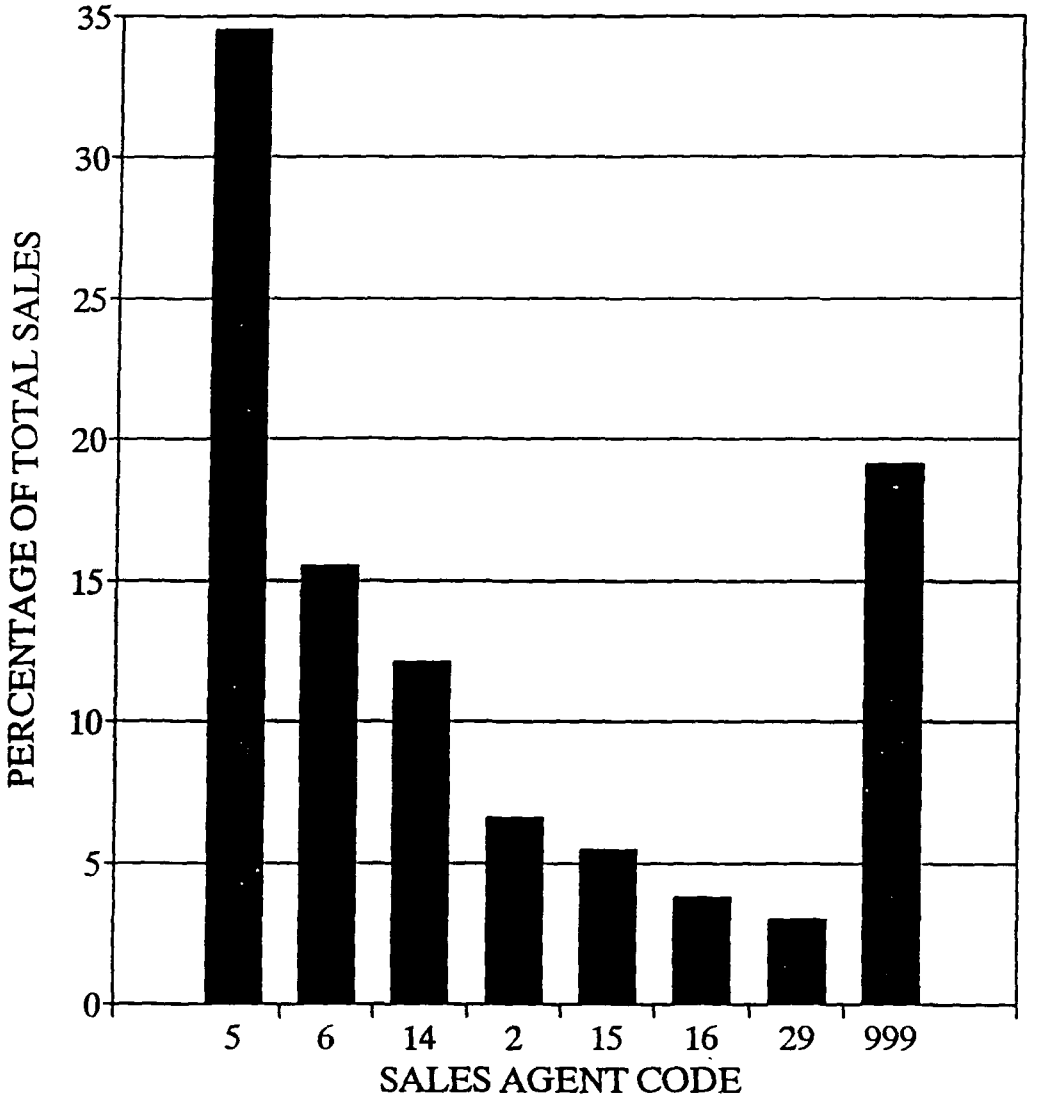

Figure 4.3.2: Histogram of Harvestore sales by sales agent code, where 999=all others.

had greater than 10% of the total adoptions. The remaining 39 sales agents each contributed less than 10% of the total adoptions. Obviously, sales agents 5, 6, and 14 were *exceptional* in terms of Harvestore sales, whilst the others were *typical* sales agents, in terms of their Harvestore sales. Defined in this manner, the original 42 sales agents were divided into two distinct groups in terms of Harvestore sales generated: SAL=1 for exceptional agents, SAL=0 otherwise. Further, this binary variable facilitated discerning the effectiveness, strategies and spatial pattern of adopters associated with both the typical agents and the exceptional agents. Note, however, the number of adoptions was the only criteria used in this binary definition, while the length of employment was not considered explicitly for any of the agents, but one could easily imagine some sales agents who accounted for very few adoptions (i.e., also classified in the typical category) as being employed for a shorter time periods.

Based on the prominent role played by sales agents in the adoption of the Harvestore, two distinct spatial patterns and, therefore, associated parameter estimates, are likely to result from the typical versus exceptional sales agent definition. Given the selling strategy of sales agents and their use of the neighbourhood effect as revealed in Scorgie (1973; 1992), it is hypothesized that most sales agents likely remained in smaller, well-defined regions of the study area. If this pattern is true for the typical sales agents, this would be reflected in positive parameter estimates for the SAL variable. More specifically, a significant positive parameter estimate for SAL would indicate that longer durations are associated with Harvestore adopter locations sold by an exceptional sales agent and shorter durations exist at adopter locations sold by the typical sales agents.

Conversely, a significant negative parameter estimate would indicate that adopter locations associated with an exceptional sales agent tend to be in closer proximity to other adopters (i.e., shorter durations) than adopter locations associated with typical sales agents are farther apart (i.e., longer durations), relatively speaking. If this second hypothesis holds true, then the exceptional sales agents may be associated with a few large clusters of adopter locations or many small clusters of adopters.

Similarly, other independent variables connected with the interpersonal sources of communication and spatial mechanisms of information flow, based on the notion of *homophilus individuals* could also effect the length of durations associated with adopter locations. In this regard, the type of farming operation (hog, beef or dairy) and the feed type used (corn, haylage or grain) may have influenced the adoption of the Harvestore in the study area. Intuitively, farmers with similar farm operations and using similar feed types will likely encounter similar circumstances in day to day operations and the communication level between them is most probably high. Thus, the adoption or rejection of sealed storage and the Harvestore amongst farmers would likely be influenced by the opinions of homophilus individuals interacting. In terms of measuring this effect in an event history methodology, one must consider the possibility that shorter or longer durations are associated with particular farm and feed types.

The percentage of adopters by farm type is shown in Figure 4.3.3 where "other" indicates some multi-type farms. The predominant farm types are dairy (44.3%) and hog (36.0%), whilst beef (6.6%) and other (13.1%) are much less common amongst adopters. In terms of feed type, Figure 4.3.4 clearly shows that the majority of adopters

of the Harvestore utilize the sealed structure for the storage of high moisture content corn. This is not surprising given the preceeding discussion in section 4.2. The second most popular feed type used by Harvestore adopters is haylage, again the high moisture content of this feed type makes sealed storage appropriate.

Given this situation, a set of dummy variables for farm type (COR, HAY) and a set for feed type (DAI, HOG) are considered in the regressive event history analysis. The hypothesis for the farm and feed type dummy variables are that stronger communication ties might exist amongst farmers with the same type of farm operation or those using the same feed type. To some extent, the variation in the spacing of adopter locations based on communication by homophilus individuals can be measured to determine the degree in which homophilus farmers (in terms of farm and feed type) act as contagions in the diffusion of the Harvestore. A significant positive parameter estimate for one of the farm types indicates longer durations associated with this farm type relative to durations associated the alternative farm types. Likewise, any significant positive parameter estimate for one of the feed types indicates longer durations relative to durations associated with the alternate feed types. Clearly, one would expect to see a negative parameter estimate for the COR dummy variable relative to all the other feed types, since the Harvestore is ideal for high-moisture content corn and farmers using this feed type might share their knowledge of the Harvestore to neighbouring farmers using corn feed. However, it is important to note that the farm and feed type data assumes that the spatial variation in farm and feed types is homogeneous. This is an important assumption to consider, especially in the context of modelling durations

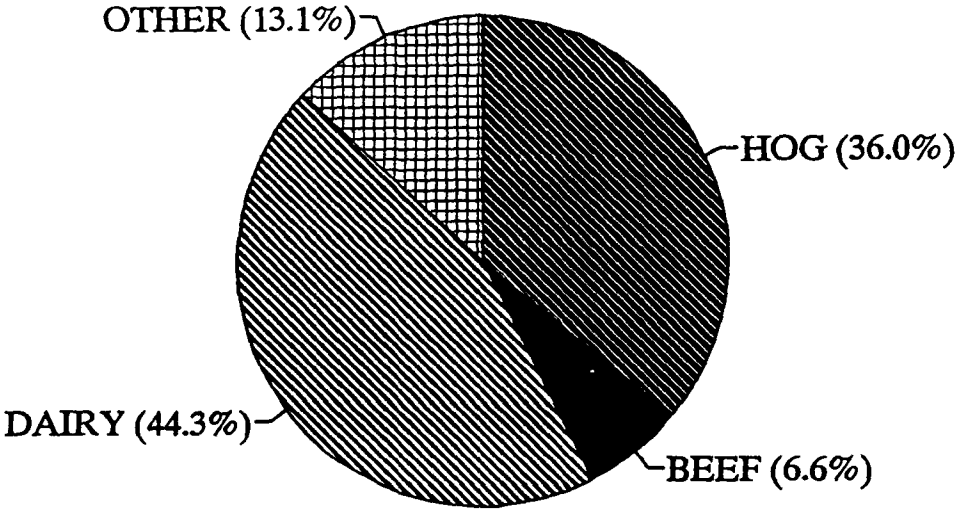

Figure 4.3.3: Percentage of adopters by farm type.

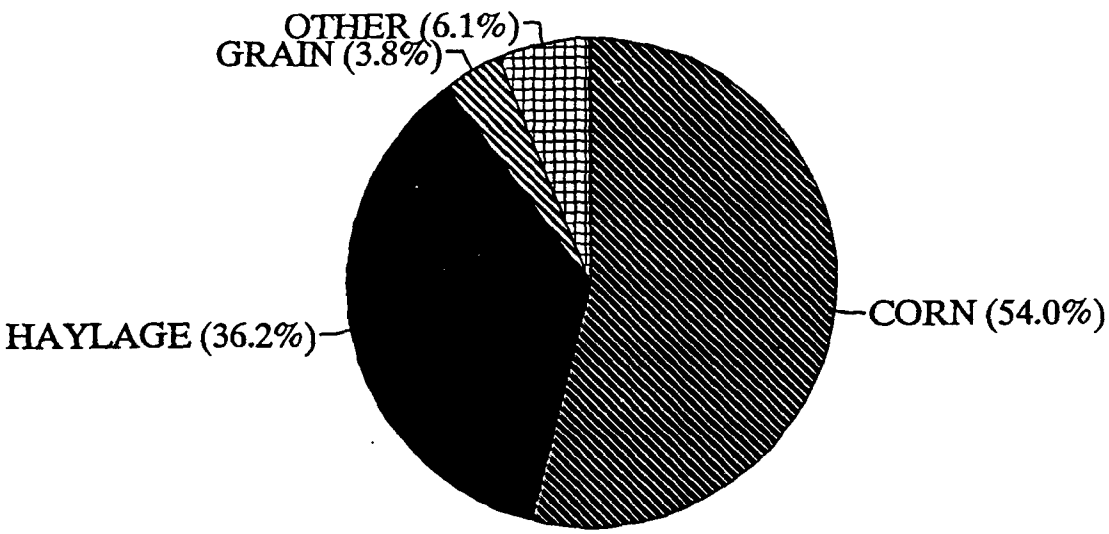

Figure 4.3.4: Percentage of adopters by feed type.

between events, where any spatial variation amongst farm and feed types may well obscure the effects of homophilus farmers on duration.

In contrast to the preceeding explanatory variables that were concerned with mechanisms of spatial information flow, variables regarding the spatial structure of the study area are now defined. In this respect, these variables attempt to measure environmental heterogeneity of the study area. The importance of accounting for as much measurable environmental heterogeneity as possible was discussed in Chapter Three and is important with regard to identifying duration dependencies in the data as opposed to spurious duration dependencies caused by unobserved or unmeasured environmental heterogeneity. A definite factor affecting the distribution of Harvestores in the study area is the number of commercial farmers in each are. Therefore, the *density* (DEN) of commercial farming operators in each township was calculated within the TransCAD GIS database using census figures from 1971, 1976 and 1981 for consistency. Each adopter location was associated with the farming density measure of the township in which it is located. This information is shown in Table 4.3.1. Accounting for variations in farming density may explain areas with inherently shorter durations between adopter locations. One would suspect the DEN variable to have a significant negative parameter estimate which would imply that areas of higher farming density account for shorter durations, associated with adopter farms in those areas, as compared to the durations associated with adopter farms in the remaining, lower density areas.

The final explanatory variable in this regression analysis also concerns the spatial structure of the study area. A *Corn Heat Unit* (CHU) is defined as the number of days

COUNTY	TOWNSHIP	1971	1976	1981	AREA +	DEN + +
Brant	Dumfries, S.	271	236	262	71.62	3.58
Huron	Grey	327	276	291	109.80	2.71
	Howick	401	349	325	114.20	3.14
	Hullett	300	267	290	87.36	3.27
	McKillop	292	255	288	90.16	3.09
	Morris	292	254	244	92.69	2.84
	Tuckersmith	238	196	221	69.56	3.14
	Turnberry	175	150	160	60.89	2.66
	Usborne	259	239	245	70.88	3.49
	Wawanosh, E.	203	186	174	69.08	2.72
Middlesex	Biddulph	241	212	219	64.89	3.45
	London	512	430	378	135.85	3.24
	Nissouri, W.	318	283	263	86.38	3.33
Oxford	Blandford	156	441	485	47.12	3.10
	Blenheim *	363			108.44	
	Nissouri, E.	309	690	679	80.63	5.16
	Oxford, N.*	142			36.98	
	Zorra, E.	427	383	380	96.42	2.68
	Zorra, W.*	348			94.81	
Perth	Blanshard	301	258	269	81.81	3.37
	Downie	323	262	266	83.41	3.40
	Easthope, N.	294	279	284	70.78	4.04
	Easthope, S.	177	158	160	39.50	4.18
	Ellice	356	333	329	92.16	3.68
	Elma	425	394	412	113.10	3.63
	Fullarton	264	230	217	67.19	3.53
	Hibbert	233	207	194	67.74	3.12
	Logan	366	330	331	93.17	3.67
	Mornington	367	345	383	86.81	4.20
	Wallace	343	316	311	86.20	3.75
Waterloo	Dumfries, N.	179	170	209	75.08	2.48
	Waterloo	314	149	143	152.58	1.32
	Wellesley	558	494	509	109.50	4.75
	Wilmot	435	346	386	107.23	3.63
	Woolwich	449	519	576	92.81	5.55
Wellington	Arthur	360	308	343	112.89	2.99
	Guelph	195	146	152	67.27	2.44
	Maryborough	344	305	323	93.23	3.48
	Minto	393	344	357	120.20	3.03
	Nichol	186	132	152	49.23	3.18
	Peel	456	446	470	125.38	3.65
	Pilkington	175	155	161	48.86	3.35
		total	total	total	mean	mean
		13067	11473	11841	86.28	3.38

+ Calculated with TransCAD in miles squared.

+ + Density calculated with average number of operators [(1971+1976+1981)/3].

* Included with above Township after 1971.

Table 4.3.1: Farming density by township.

that the temperature exceeds 50 degrees Fahrenheit in the day and 40 degrees Fahrenheit in the evening during a growing season based on the latest 30 year record of temperatures. In other words, CHUs are analogous to the more familiar "growing degree days" used to classify geographic areas by length of growing season. In southern Ontario, the minimum CHU required for corn to mature is 2000; the entire study area exceeds this minimum. However, the level of CHU's has a definite north - south trend in the study area where the level of CHUs increases steadily when moving from the north to the south of the study area. In this regard, we can hypothesize that farmers in areas with lower CHUs may be more apt to adopt sealed storage to preserve their smaller yield as opposed to farmers in the south (higher CHU's) who likely have less need for a Harvestore due to the availability of corn for longer time periods (Scorgie, 1992). If this were true, the event density would increase with decreasing CHU, and this spatial variation of events would result in a significant negative parameter estimate. However, a significant positive parameter estimate is also reasonable to hypothesize. In this respect, the farmers in areas of higher CHU's may want to use all of the larger corn supply they grow to increase farm operations.

From the discussion above, it is apparent that the spatial structure and information flow explanatory variables may either increase or decrease the duration (or distance) between adopter locations. This increase or decrease in duration length is also interpreted as a decreasing or increasing effect on the hazard rate, respectively. For some variables, it is reasonable to discern contradictory hypotheses for the influence of that variable on duration length, as with the effect of CHU's. Indeed, the complex

nature of the innovation adoption process (as discussed in Chapter Two) suggests that many more variables, some difficult to quantify, would be required to completely model such a process. In any event, the approach adopted here is an original attempt to incorporate explanatory variables into, essentially, a distance based, spatial point pattern analysis and the results should provide further insights into the utility of an event history approach to modelling spatial point patterns.

4.4 The Study Area

The region chosen for this study includes 42 townships in 8 counties in southwestern Ontario and comprises approximately 3,650 square miles (9,500 square kilometres or 2,300,000 acres) in total land area. The county and township level maps of the study area are shown in Figures 4.4.1 and 4.4.2, respectively. The study area was selected primarily because of the availability of detailed Harvestore diffusion data collected in Scorgie (1992). This area is appropriate for a study of innovation diffusion since southwestern Ontario is located in the recently expanded North American corn belt - where the development of hybrid corns and their storage has extended the geographical limits of commercial farming as previously discussed.

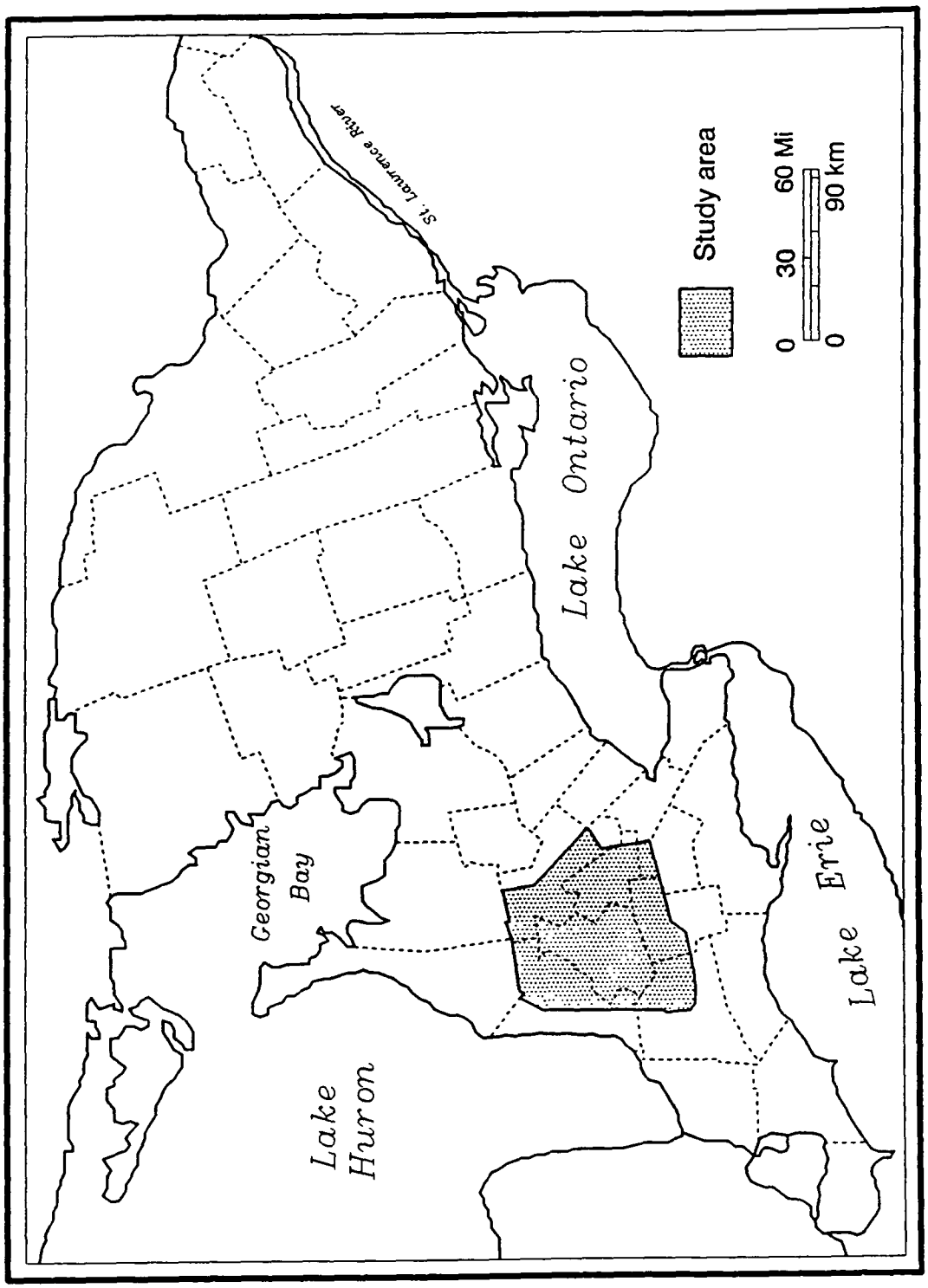

Figure 4.4.1: County level map of the study area.

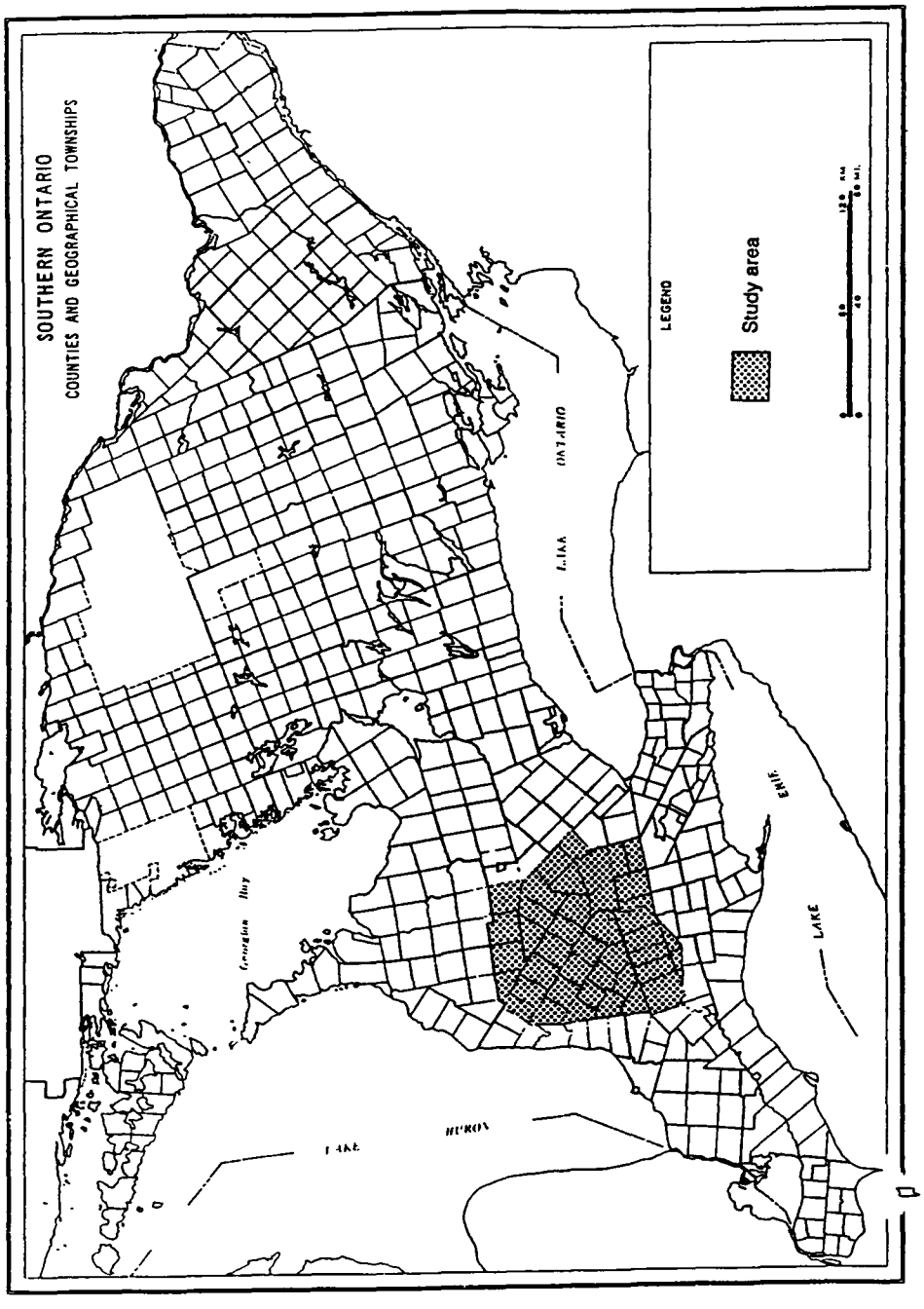

Figure 4.4.2: Township level map of the study area.

4.5 Outline of Analysis

Typically, the analysis of an event history data set commences with an exploratory stage in order to identify potential explanatory variables that correlate with the behavior under investigation (i.e., migration, shopping). In the context of this thesis, the data set is a planar point pattern and as such, the exploratory phase of analysis is concerned with comparing the empirical point pattern to a model of Complete Spatial Randomness (CSR) defined fully in Chapter Two. Consequently, the first part of the analysis utilizes powerful graphical tests and simulated spatial point patterns to explore the data and determine whether or not the set of events, taken as a whole, should be considered as randomly distributed or, in some way, clustered or regularly distributed.

To review, the hypothesis of CSR implies two important assumptions. First, that the intensity (mean number of events per unit area) does not vary over the plane. Second, that there are no interactions amongst the events. In other words, there is an assumption of independence which would be violated if an event at a location either encouraged or inhibited the occurrence of other events in that neighborhood. Basically, CSR represents an idealized standard which, although nearly always untenable in practice, is a valuable aid to the formulation of hypotheses concerning pattern and process. Thus, most spatial point pattern investigations start with a test of CSR. Diggle (1983) outlines several reasons for this approach. First, a pattern for which CSR is not rejected hardly merits any further formal analysis. Second, tests of CSR are a means of exploring the data set; the rejection of CSR is not particularly important *per se*. Finally, CSR acts as a dividing hypothesis to distinguish between patterns which include "regular"

and "clustered" at either end of the spectrum (see Chapter Two).

Rejection of CSR via graphical tests leads to the second stage of analysis concerned with the statistical evaluation of a series of explanatory variables using an event history modelling approach. The regression analysis involves modelling the hazard function (which describes the process) as dependent on the duration lengths and different explanatory variables. Regression methods permit the consideration of the influence of the spatial mechanisms of information flow and spatial structure on the pattern of Harvestore adoptions. Two regressive methods are employed in this thesis: the partially-parametric Cox model and the fully parametric models. Recall from Chapter Three that the Cox model can determine the influence of explanatory variables without having to make further assumptions about the form of the hazard rate. In contrast, fully parametric models are used when certain hypotheses concerning the duration dependency are to be tested.

4.5.1 Graphical Analysis

The analysis begins with graphical tests designed to check for randomness in the spatial point pattern. This analysis involves the construction of a theoretical distribution to represent a completely spatial random (CSR) pattern in an area of the same shape and size as the study area with the same number of points from computer simulation. The test proceeds by computing an empirical distribution function (EDF) which is the observed proportion of nearest neighbour distances. The EDF is plotted against the theoretical distribution function from the simulation. If the data are compatible with

CSR, the graphical EDF plot should be roughly linear. The theoretical distribution function can also be derived by mathematical approximation, rather than simulation, and this avenue is also explored at this stage of analysis. The analysis in this thesis involves only the nearest neighbour measurements and interpretations of EDF plots are discussed in detail in Chapter Five.

At this point, several issues are worth noting regarding the graphical analysis. First, the strength of the EDF graphical test cannot be underestimated. Diggle (1983) suggests that they often make further testing of the spatial point pattern unnecessary and are almost always informative. Indeed, Diggle (1983) goes on to say that no single test statistic should be allowed to over-ride a critical inspection of the EDF plot. Second, empirical and theoretical distribution functions are derived for both data sets: the censored and uncensored durations. In this regard, the effect of censoring is considered at every stage of the analysis, including the regressive component described below. Lastly, reciprocal nearest neighbours, although often included in statistics describing spatial point patterns, are excluded in the evaluation of the empirical and theoretical distribution function plots. This does not effect the results so long as the chances of reciprocity are independent of the lengths of the durations, and acts as a precaution against the potential problem of "dumbbells" affecting pattern inference discussed in Chapter Two. However, reciprocal durations are not a concern in the event history regressive analysis, where the focus shifts from the general event pattern to duration dependency and the effects of explanatory variables on all durations.

4.5.2 Event History Regression Modelling

Initially, the analysis of the spatial durations derived from the nearest neighbor measurements between adopter locations is concerned with the fit of parametric models (see Chapter Three). Here, the fit of the various parametric forms is of intrinsic interest without including the explanatory variables defined earlier in this chapter. Parameter estimates for the exponential, Weibull, Gompertz and log-logistic models are derived and the associated hazard rate curves are produced and evaluated.

This stage of analysis re-introduces the reciprocal durations so that each adopter location's duration could be associated with a series of explanatory variables. The regression analysis is initially concerned with examining the effects of all the explanatory variables. Subsequent model specifications remove variables not considered to have a significant effect on model fit. In all cases (i.e., either no variables or various regressive specifications), the main interest was the fit of various parametric models and the effects of censoring. The estimated hazard rate curves calculated at the mean of the explanatory variables are plotted to provide insight into the distributional form of spatial duration dependence. The final stage of the regressive analysis involves using the partially parametric Cox model to model the durations, and determine, in an unbiased manner, the effects of explanatory variables on the durations. Again, in all cases, both the censored and uncensoring data sets are examined independently and the results compared.

4.6 Description of Applied Software

Fortunately, several powerful software packages for the analysis of event history data exist and one in particular, LIMDEP, provides all of the features and flexibility required by this research. LIMDEP (Greene, 1990) is a flexible econometric software package which can be used to estimate a wide variety of models including the continuous time event history models of interest in this research. Strictly command driven, LIMDEP provides all the parametric and partially-parametric estimation procedures (including maximum likelihood). In addition, LIMDEP allows for macro programming in the form of user written procedures which were indispensable to this research when the simulation results required analysis.

The other major software utilized in this research included two geographical information system (GIS) packages. The first, TransCAD, was developed by Caliper Corporation and can be described as a menu-driven, PC-based, full featured GIS. TransCAD provides a database manager with tools such as spreadsheets, pie and bar charts, and statistical analysis routines. In addition, TransCAD incorporates a set of network analysis, operations research and transportation models suitable for many applications. In the context of this thesis, TransCAD's powerful map display and data manipulation tools were utilized to derive summary statistics of the data and to calculate farming densities per township.

The second GIS was IDRISI, a grid or raster based, low cost, menu-driven teaching package developed by Ronald Eastman at Clark University. IDRISI was specifically created for microcomputers using DOS that allows users to analyze and

manipulate spatially referenced data using a series of independent modules linked by a simple data structure (Eastman, 1988; 1991). A raster GIS divides space into a grid of cells for which various attributes (e.g., land use, altitude) may be stored. This regular structure makes IDRISI particularly appropriate for coupling with high level programming languages using similar logical data structures. For this research, IDRISI's grid referencing system provided the information needed to reconstruct the study area shape and size for the simulation routines. Specialized programs for simulating the point process and data conversion were written in Fortran by Dr. Steven Reader and the author, while custom estimation routines for LIMDEP were written by Dr. Reader. Additional graphical output was achieved using AXUM and Quattro Pro software packages.

IT IS BETTER TO KNOW SOME OF THE QUESTIONS THAN ALL OF THE ANSWERS.

JAMES THURBER

CHAPTER FIVE

RESEARCH FINDINGS AND DISCUSSION

5.1 Tests of Complete Spatial Randomness

The first stage of the data analysis required the formulation of a computer program to generate 100 simulated CSR spatial point patterns. This simulation exercise provided the experimental (or theoretical) distribution of nearest neighbour durations from a CSR process in a bounded region identical to the empirical study area. This simulation data was then used in a graphical evaluation of the empirical durations versus durations generated from the CSR process. Having some knowledge of the characteristics of the event pattern aided in interpreting the explanatory variable coefficients and the parametric models that best fit the durations. For example, if the empirical durations showed clustering versus the experimental CSR durations, then the regressive analysis could focus on the possible clustering mechanisms (i.e., perhaps the contagion(s) or environmental heterogeneity). Hence, it was of intrinsic value to the interpretation of parameter estimates affecting duration lengths whether or not the observed event pattern was clustered, random or regular. The logic of the computer simulation program is described below, and Appendix One contains the actual Fortran

77 code.

The simulation program proceeded by randomly generating 528 coordinate locations (i.e., events) in Cartesian space. Several obstacles had to be overcome in accomplishing this task, however. For instance, the simulated events had to be contained within the same size (to scale) and irregularly shaped study area as the empirical data. This problem was handled by including a sub-routine in the program that evaluated each XY coordinate location to determine its inclusion or exclusion from the simulated study area using an interface with IDRISI, the grid referencing (or raster based) geographical information system (GIS) software package described in Chapter Four. A cartographic modelling or map algebra approach was used, applying Boolean logic MAP OVERLAY functions to isolate the study area from surrounding (or external) map pixels. Cartographic modelling involves using the output from one analytic mapping function as the input data (i.e., map) for the next function (Arnoff, 1989). In addition, the RECLASS function enabled the distinguishing of coordinate locations (i.e., events) within the study area, from those on its boundary, as well as from any external coordinate locations. The digitized map stored in IDRISI was recoded so that the internal coordinate locations of the study area had a value of 1 associated with them. Similarly, the boundary coordinate locations were recoded to 2, whilst the external coordinate locations were coded 0. By comparing the simulated event coordinates to the IDRISI recoded coordinate locations, the simulation program could determine inclusion or exclusion of each randomly generated coordinate location. As an aside, the IDRISI GIS analysis also provided measures of the study area perimeter and area which were used

as input variables to approximation formulae used later in the analysis.

The simulation process was repeated 100 times so that an experimental random distribution of nearest neighbour durations could be derived and used as a CSR standard to compare to the empirical durations. The necessity of this experimental distribution was based on the fact that approximation formulae for nearest neighbour durations under CSR are typically derived based on the assumption of an unbounded plane and thus do not account for edge effects. Alternatively, if the formulae are derived for a bounded region, they are unreliable for use with highly irregularly shaped regions, as is the case with the empirical study region (Diggle, 1983). An experimental distribution provided a theoretically attractive distribution of nearest neighbour duration measurements under the conditions of CSR within a bounded region. Therefore, we could be reasonably confident of any pattern inferences made concerning the empirical distribution of nearest neighbour durations as compared to durations measured from the 100 random simulations. It is also important to note that experimental distributions were derived for both the uncensored case, and the case in which some durations were censored. This approach facilitated an evaluation of the utility of measuring spatially censored durations.

After generating the simulated random spatial point patterns, the program computed nearest neighbour distances (or durations) between events. To achieve this, a *crude search algorithm* was used to explore all the inter-event distances from each event until its nearest neighbour event was found. Alternative algorithms for deriving nearest neighbour durations exist, however, including the construction of Dirichlet tessellations around the n events and then searching for the nearest neighbour distances

within the tessellation (i.e., the *Green-Sibson algorithm*). Essentially, this tessellation method makes use of the fact that each event is contiguous to, on average, about six other events, one of which must be its nearest neighbour. Consequently, only a small fraction of the inter-event distances need to be calculated resulting in decreased computational costs. However, Diggle (1983) notes that the crude search method is more efficient than the tessellation method for n less than 500, but thereafter becomes progressively less efficient with increasing n. Hence, the crude method was adopted in this thesis because the relative gain in computational efficiency for $n=528$ was trivial in comparison to the increased programming complexity.

In addition, the simulation program was purposefully designed in an adaptable format to achieve two main goals. First, the program had to check for reciprocal nearest neighbours and "flag" these durations so they could be eliminated as required. Removing one reciprocal duration from each pair eliminated the chance of spurious pattern inference associated with the existence of "dumbbells" of events which could act to obscure event pattern evaluations when using nearest neighbour distances. For instance, in Chapter Two it was noted that a spatial point pattern with a substantial amount of dumbbell events, and therefore reciprocal durations, could actually be statistically classified as all three general pattern types, dependent only the distribution of the dumbbells, whilst visually, the event pattern actually resembled an alternative pattern type. By eliminating the reciprocal nearest neighbours at this stage of analysis, the problems associated with pattern inference from dumbbells of events were not an issue. Although removing these reciprocal durations decreased the number of durations to

analyze, it maintained methodological consistency, so long as the chances of reciprocity were independent of the lengths of durations; no theoretical basis for rejecting this assumption exists in this analysis (Clark and Evans, 1954; Odland and Ellis, 1992).

The second goal achieved by programming flexibility was the derivation of both a *censored* and an *uncensored* experimental CSR distribution to compare to the empirical distribution of nearest neighbour durations. For the uncensored durations, the program computed the traditional nearest neighbour durations for each of the 100 simulated data sets. In the censored duration case, however, the program's crude search for a nearest neighbour event included exploring the distances from each event to each coordinate location on the boundary resulting in two possible outcomes: a regular nearest neighbour duration terminating at an event (equivalent to an *uncensored* duration), or alternatively, a *censored* duration from an event (i.e., a boundary event) which terminated at the nearest boundary coordinate location. In other words, the censored observation indicated that, for a certain boundary event, no other event was closer to this event then the distance from the boundary event to the boundary coordinate.

As expected, the removal of reciprocal durations resulted in a decrease in the number of durations to be examined. On the other hand, a slight increase in the number of durations occurred by allowing durations to be censored. This result is illustrated in the form of a histogram in Figure 5.1.1 where the censored simulations were shown to routinely contain, on average, 17 more durations than the uncensored simulated data sets. This result was easy to explain intuitively since the actual censored observations, within the simulations allowing for censoring, could not have a reciprocal, and consequently,

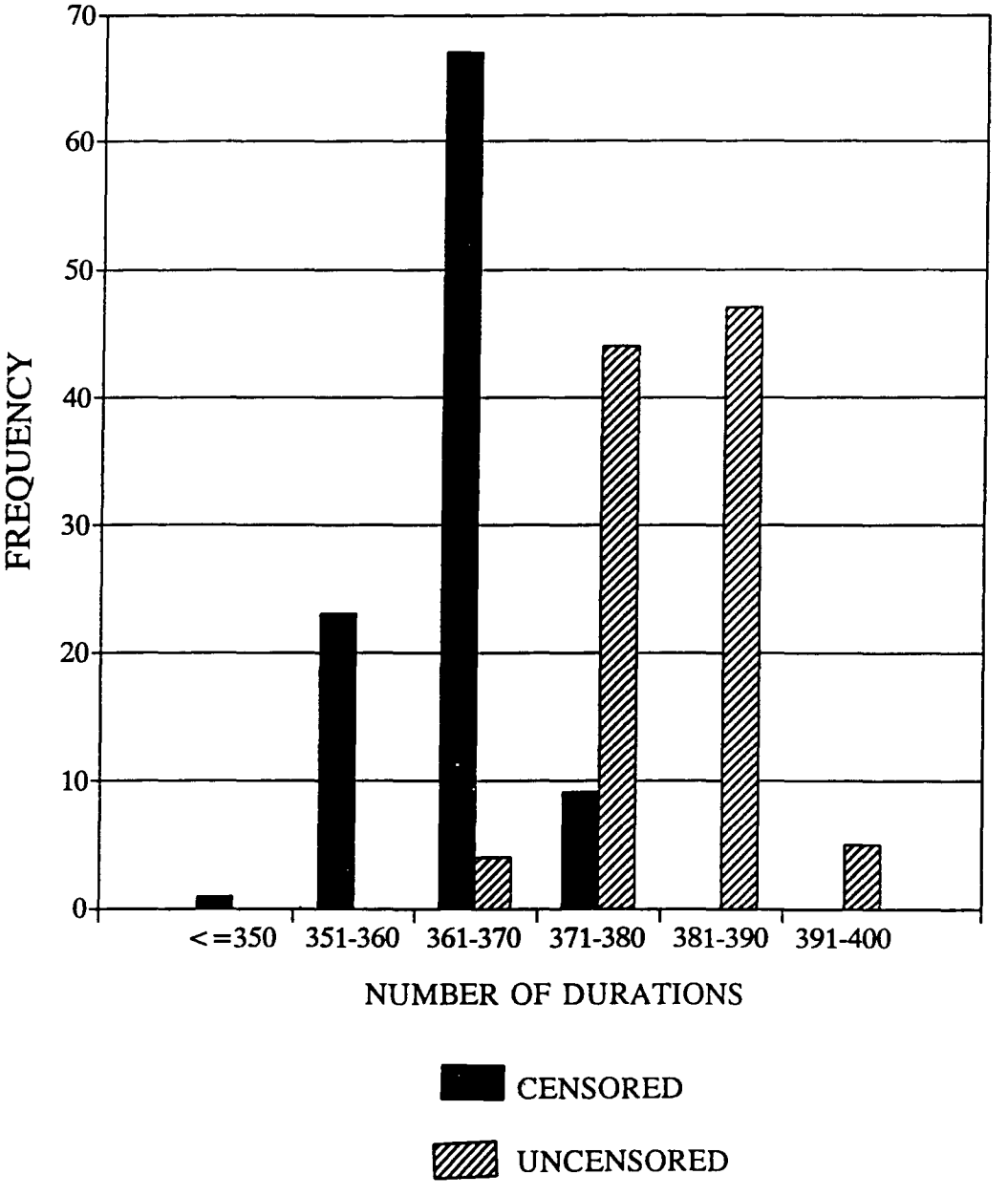

Figure 5.1.1: Histogram of number of durations in censored and uncensored simulations.

would contain more durations than the uncensored durations which have had the reciprocal durations removed. The average number of durations minus the reciprocals for the censored data sets was 380.5 (range 350 to 376), whilst the uncensored data set averages 363.5 durations (range 364 to 397) per simulated spatial point pattern. When the portion of the simulation program for computing nearest neighbour durations was run on the real data, the uncensored version contained 374 durations, while the censored version had 387. Again, these nearest neighbour durations were minus the reciprocals, and the numbers were consistent with those in the simulation exercise.

It was mentioned earlier that approximation formulas to test for CSR were available, although subject to limitations. At this point, it was interesting to contrast the simulation CSR duration results with a test of CSR derived from approximation formulae proposed by Donnelly (1978) for the normally distributed sample mean and variance of nearest neighbour durations under CSR; formally,

$$E(\bar{x}) = \frac{1}{2}(n^{-1}|A|)^{\frac{1}{2}} + (0.051 + 0.042n^{-\frac{1}{2}})n^{-1}\, l(A) \qquad (5.1.1)$$

$$Var(\bar{x}) = 0.070n^{-2}|A| + 0.037(n^{-5}|A|)^{\frac{1}{2}} l(A)\ , \qquad (5.1.2)$$

where $l(A)$ denotes the length of the boundary of a region with area A and n events (here we used the perimeter and area calculations from the earlier Cartographic Modelling). Significantly small or large values of x would indicate aggregation or regularity, respectively. However, in very convoluted regions A, this approximation often breaks down and implementation of graphical tests based on experimental distributions is

recommended (Diggle, 1983). By comparing the estimates derived by the approximation formula from the estimates derived by simulation, we can assess the validity of these formulae with regards to the empirical data used in this thesis. Since the empirical study area is irregularly shaped, we might expect the approximation formulae to be unreliable.

When the area, perimeter and number of events for the empirical data set were input in equations (5.1.1) and (5.1.2), they yielded an expected sample mean of 8.518 with variance 0.427 for nearest neighbour durations, if the spatial point pattern were, in fact, CSR. To compare these approximation figures to the simulated CSR patterns, the mean of the means of all the 100 simulated data sets (for both the uncensored and censored cases) had to be determined, and then the variance around this mean of means was calculated. The simulated durations have mean 8.569 and variance 0.054 for the data sets with censored durations, and mean 9.598 and variance 0.037 for the uncensored durations. In the case of the censored data, the approximation formulae and the simulated random process results were not significantly different, indicating a random process. However, in the case of the uncensored data, which were basically traditional nearest neighbour distances, the expected mean of the simulated durations was significantly greater than that derived from the formula, indicating regularity in the spacing of events. In spite of the simulated censored durations result, only the result of the simulated uncensored durations was directly comparable since censored measurements were not accommodated in the approximation formulae. In this respect, the Donnelly (1978) formulations incorrectly classify the simulated CSR durations as coming from a regular pattern rather than a CSR pattern. It can be concluded, therefore, that the

formulations taken from Donnelly (1978) were sensitive to the irregular shape of the empirical study area. This result underscored the importance of simulation experiments in spatial point pattern analyses given the limitations of various approximation formulae, especially in the presence of edge effects.

5.1.2 Graphical Analysis

The spatial point pattern could be formally tested for CSR with Empirical Distribution Function (EDF) plots produced by graphing the empirical distribution function of nearest neighbour durations, $R(x)$, against the theoretical distribution function of durations from CSR, $T(x)$. In this thesis, the $T(x)$ of nearest neighbour durations under CSR was derived from the experimental distribution based on 100 random simulations, for both censored and uncensored sets of durations. In this sense, the distribution of nearest neighbour durations which represented a CSR pattern in an area of the same shape and size as the study area with the same number of events provided strong evidence for or against CSR when compared with the actual distribution of nearest neighbour durations. Further, both the censored and uncensored durations were compared to their respective CSR durations derived from the simulations in order to assess the affect of measuring censored durations to account for edge effects.

The EDF plots were formed by the following procedure. Recalling the statistical concepts of event history modelling outlined in Chapter Three, the notion of a hazard rate was introduced as the dependent variable in event history model specification, along with its relationship to the survivor and density functions [see equations (3.2.7) to

(3.2.11)]. Another function related to the hazard is the cumulative or integrated hazard function, defined as

$$\Lambda(x) = \int_0^x h(u)d(u) \ . \tag{5.1.3}$$

The integrated hazard is a useful tool for model specification checks (i.e., it provides a smooth plot used to evaluate the appropriateness of various parametric models to a particular data set) and is equal to the negative of the logarithm of the survival rate (Kiefer, 1988). For the purposes of this thesis, however, we note that the distribution function of durations, F(x), may be derived from the integrated hazard in the following manner:

$$F(x) = 1 - \exp(-\Lambda(x)) \tag{5.1.4}$$

Recall that the variable x represents the length of a spatial duration. Employing this rearrangement of formulas was convenient since LIMDEP could take each distinct duration measurement from the durations produced in the simulation program and, in an automated fashion, provided the integrated hazards, which were then transformed via equation (5.1.4) to develop a *T(x)* for the EDF plots, this is explained further below.

A total of 11,180 distinct durations (or distinct nearest neighbour measurements) for the uncensored case and 10,183 distinct durations for the censored case resulted from the 100 CSR simulations. These durations were analyzed using LIMDEP's "macro-like" user-defined procedure library which was able to run the custom "survival analysis" procedure that derived the integrated hazard for each distinct duration (this program is

shown in Appendix Two). The distinct duration, and subsequent integrated hazard information, was tabulated to track the average duration, maximum duration, minimum duration and the number of occurrences at **each** distinct duration measurement.

The average, maximum and minimum distinct durations were then used to define the *T(x)*, and upper *U(x)* and lower *L(x)* simulation "envelopes", respectively. The integrated hazard value which corresponded to each distinct duration was converted to the distribution function using equation (5.1.4). The graphical procedure then consisted of plotting *T(x)*, *U(x)* and *L(x)* against *R(x)*. The resulting plot illustrates the proportion of durations which are at most length x, that would be observed under CSR (by reading the horizontal axis and the dotted diagonal line), and the observed proportion of durations which are at most length x, from the empirical data (by reading the vertical axis and the solid line). It is important to note that the durations are portrayed as distribution functions, and so, by definition, the actual duration lengths are not available from the vertical or horizontal axes. If the empirical durations were consistent with CSR, the solid and dotted diagonal lines would be very close throughout their range. The upper and lower envelopes are used to aid in interpretation of the EDF plot by illustrating the extreme cases (i.e., the maximum and minimum duration lengths observed) from the CSR simulations for a given duration length. Too few short durations (a solid line below the dotted diagonal and near or slightly below the dashed lower envelope line) indicates regularity in the spacing of events, while too many short durations (a solid line above the dotted diagonal and near or slightly above the dashed upper envelope line) indicates clustering in the event pattern (Diggle, 1983).

The information on the number of occurrences of each distinct duration was required because the longer durations were much less common, by definition, when using nearest neighbour to define durations. To avoid meaningless calculations or potentially spurious results caused by the few longer distinct durations, the average for a distinct duration was only calculated if the occurrence of that distinct duration was greater than 30 in the 100 simulations. Similarly, the maximum and minimum duration values were derived only for those distinct durations with an incidence greater than 30. This minimum of 30 occurrences tends to make the upper portion of the resulting plots rather disjointed and unreliable beyond about the .95 portion of the EDF plot, but removes the chances of meaningless duration measurements from the EDF evaluation.

Figure 5.1.2 shows the EDF plot of the uncensored durations while Figure 5.1.3 shows the EDF plot for the censored durations, recall that the $R(x)$, $T(x)$, $U(x)$ and $L(x)$ for these plots are consistent in terms of using either the censored or uncensored durations. The plots look very similar and can be interpreted in the same manner. The observed proportion of durations at a given length (i.e., the solid lines), for both the censored and uncensored data, contain an excess of shorter nearest neighbour durations than the theoretical number of durations under CSR (i.e., the dotted diagonal line). Further, the solid lines on both the censored and uncensored plots exceed the upper envelope (i.e., the dashed line above the theoretical distribution) throughout the shorter range of durations. These characteristics of the EDF plots are consistent with a clustered or aggregated pattern of events (Diggle, 1983). Thus, both the censored and uncensored durations indicate that the empirical event pattern of Harvestore adopters shows a

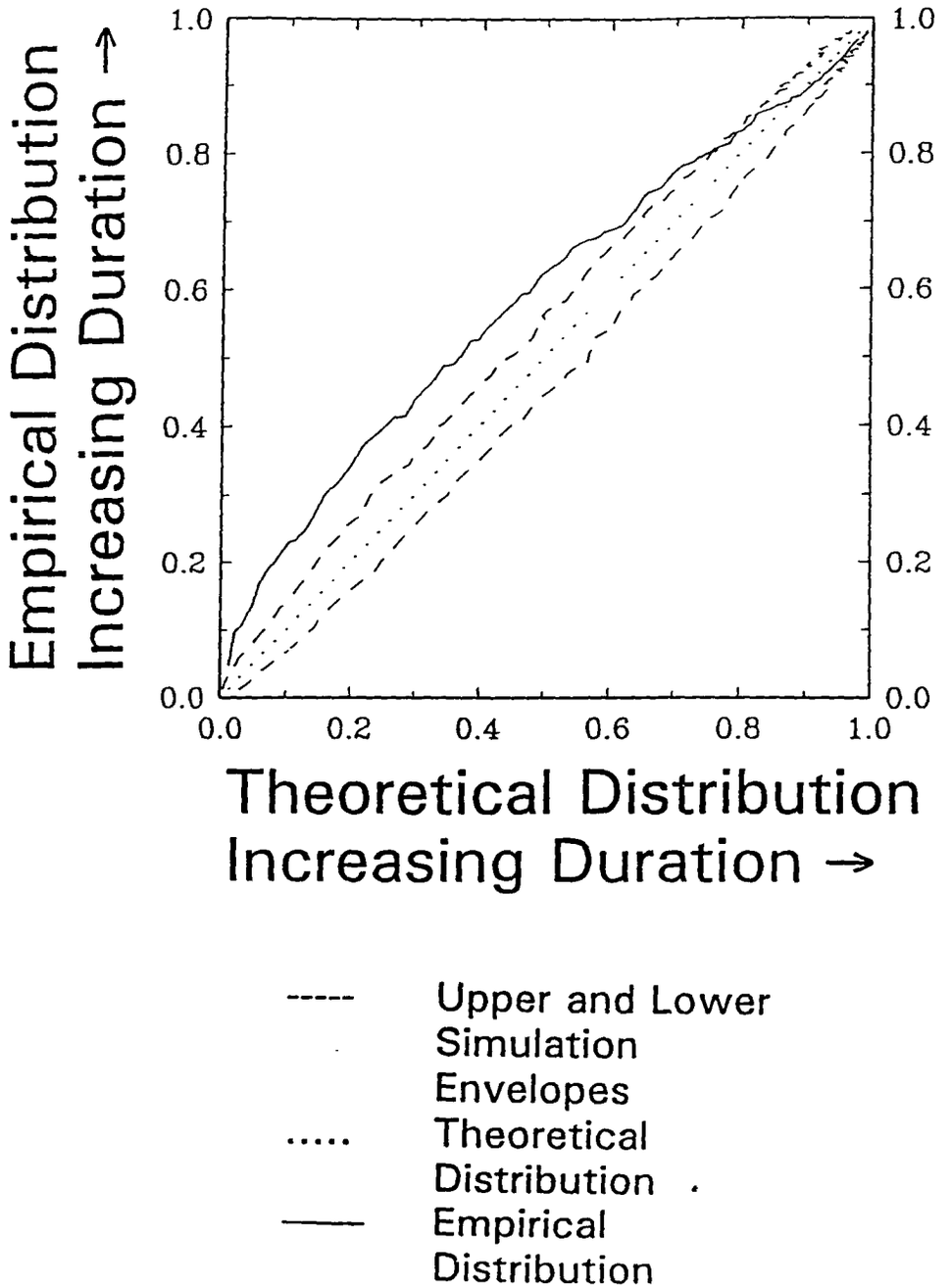

Figure 5.1.2: Empirical Distribution Plot for uncensored durations.

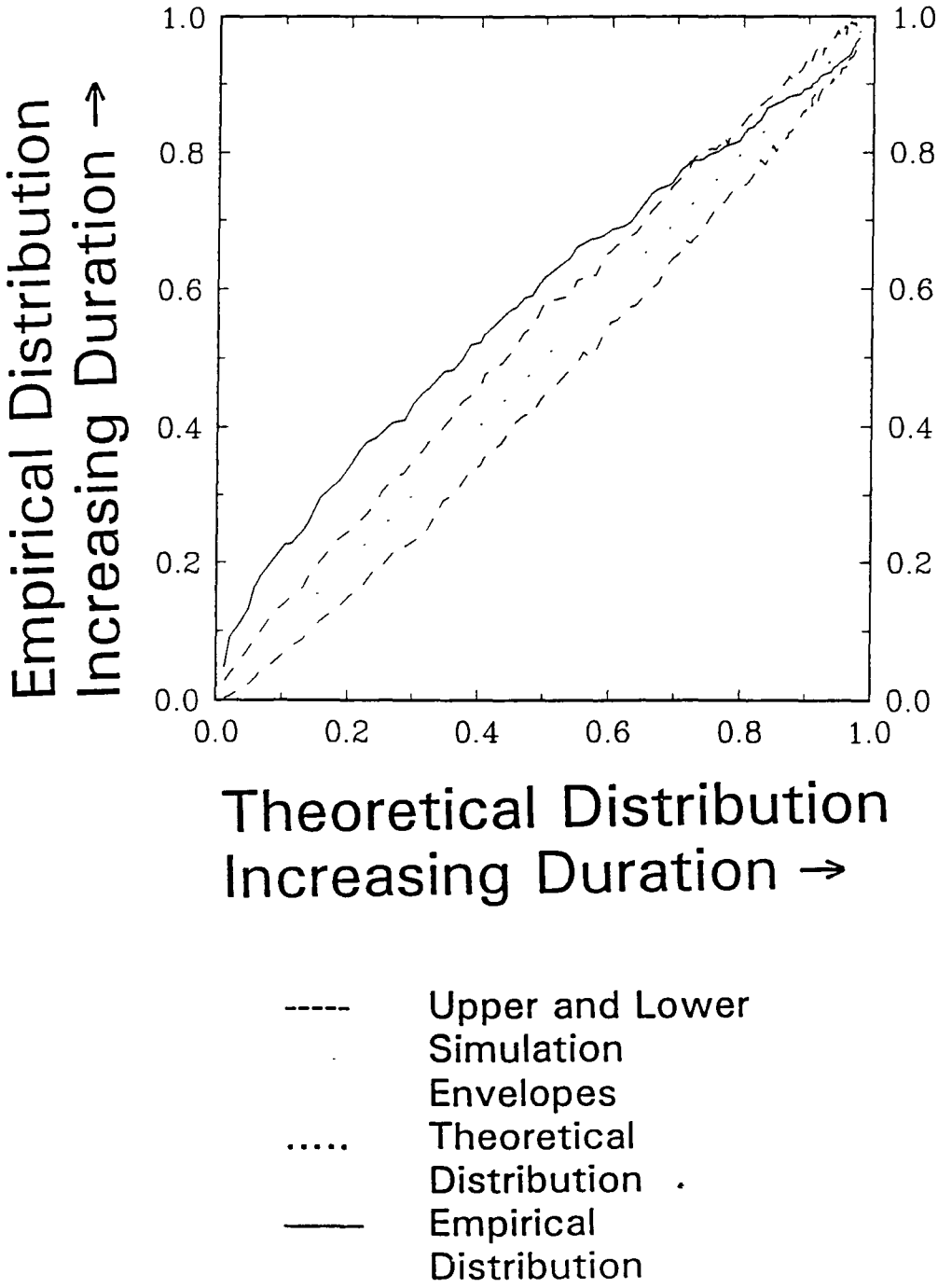

Figure 5.1.3: Empirical Distribution Plot for censored durations.

tendency towards clustering.

To be comprehensive with the analysis of the event pattern, an approximation formula for the distribution of nearest neighbour durations under CSR proposed by Diggle (1983) can also be plotted against the empirical distribution function. This formulation, however, "ignores" edge effects which means that it was derived with the assumption of an unbounded plane. Thus, it would tend to approximate a greater number of shorter nearest neighbour durations than expected in a bounded region. This being the case, this test formula could be considered a more conservative test of clustering or aggregation amongst events in space then the previous simulated CSR graphical test. The formula for the *approximate distribution function* is

$$G(x) = 1 - \exp(-\lambda \pi x^2) , \qquad (5.1.5)$$

where $\lambda = nA^{-1}$, n is the number of events, and A is the area. Figures 5.1.4 and 5.1.5 show the EDF plot of the observed distribution of uncensored and censored durations, respectively, against Diggle's (1983) approximation formula for a CSR distribution. Once again, both the censored and uncensored durations showed similar patterns on the graphs and can be interpreted in the same manner. The observed durations (i.e., the solid lines) illustrate the excess of shorter durations then would have been expected under CSR according to the approximation formula. Again, this is an indication of a clustered pattern of events. The solid lines tend to follow the upper simulation envelopes (i.e., the upper dashed lines) throughout their range, as opposed to exceeding the upper envelope in the theoretical EDF plot earlier. This indicates that the approximation formula did, in fact, predict a larger proportion of shorter durations since it "ignores" edge effects,

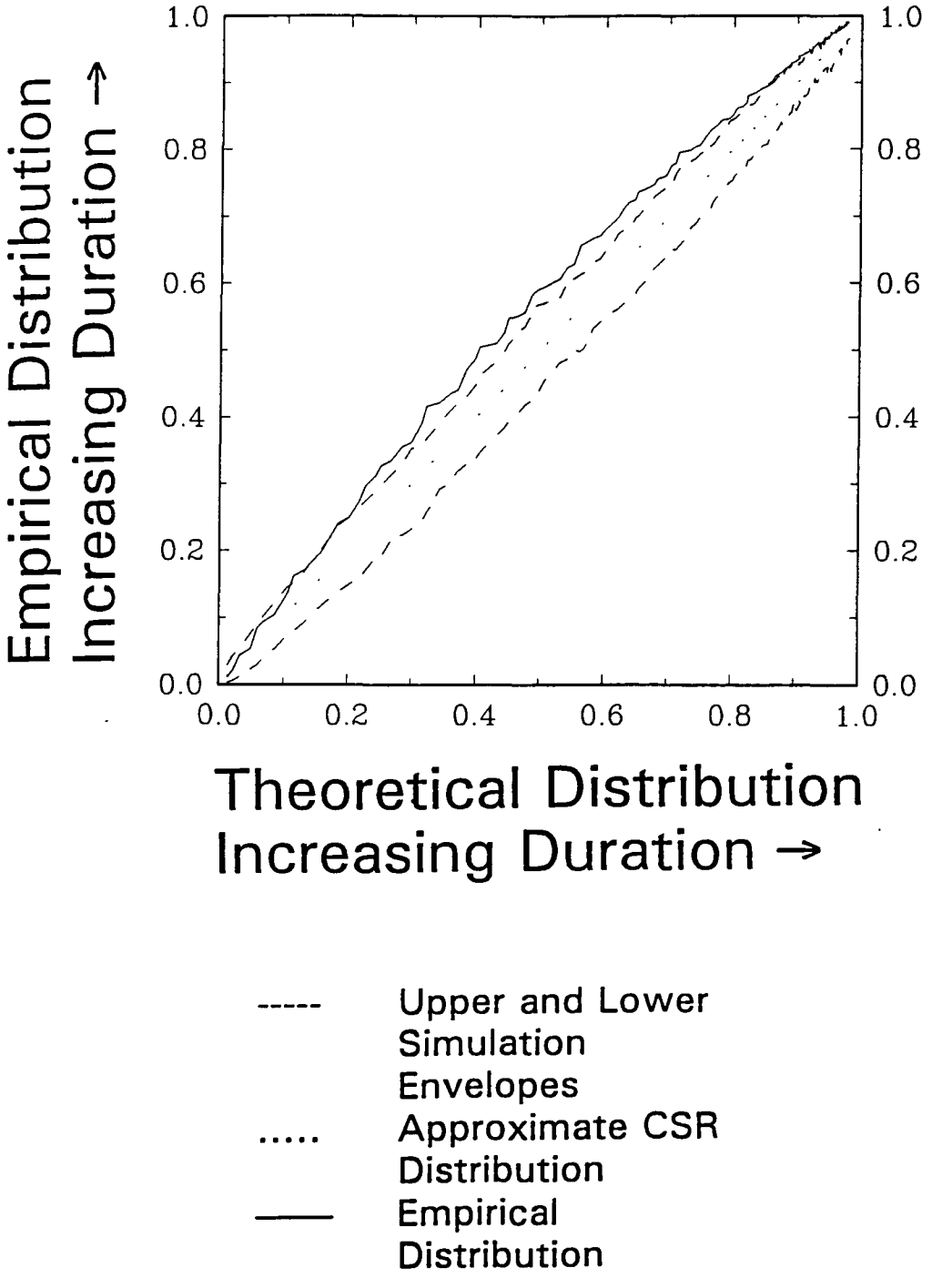

Figure 5.1.4: EDF Plot using Diggle's Approximation with uncensored durations.

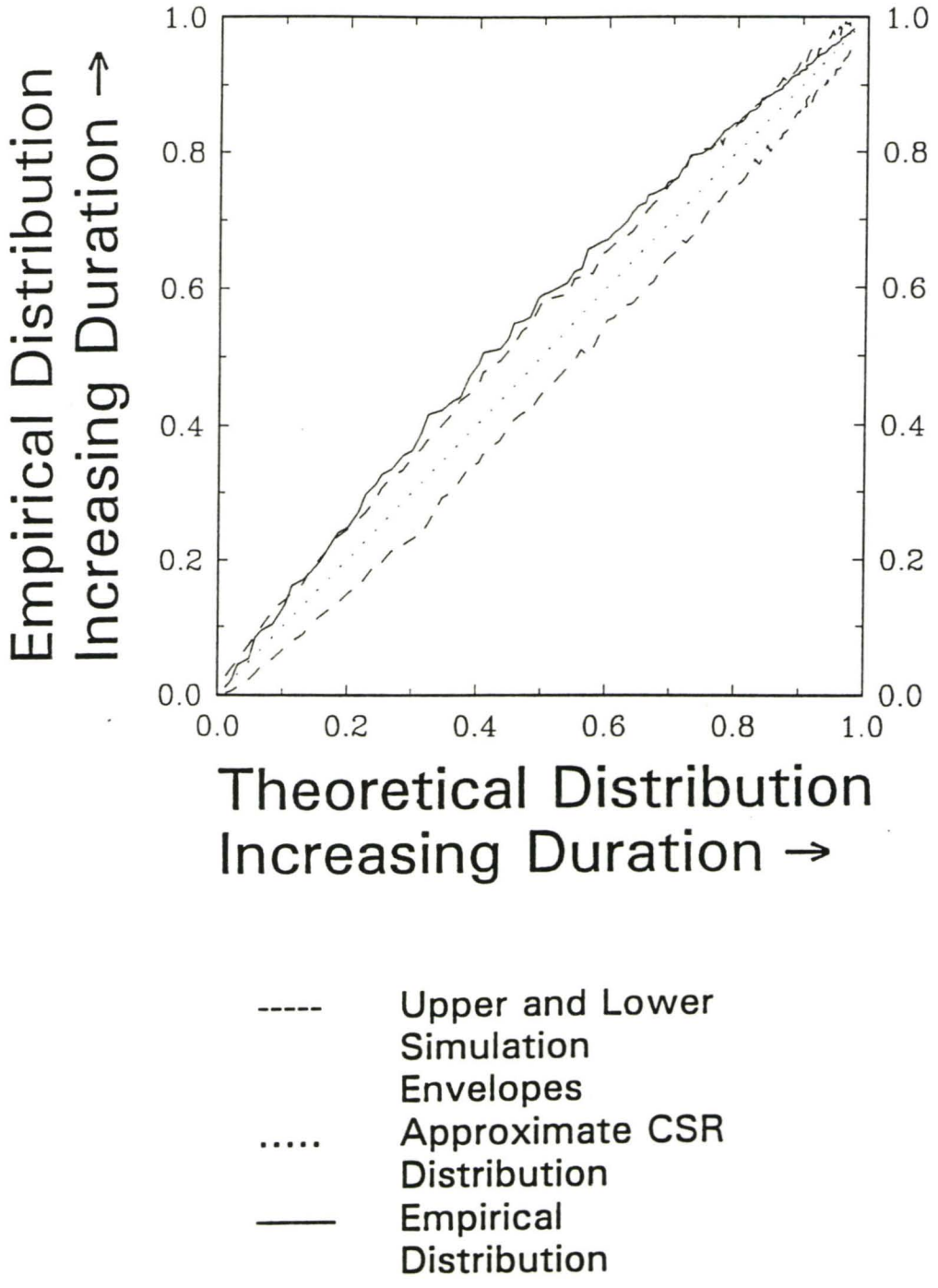

Figure 5.1.5: EDF Plot using Diggle's Approximation with censored durations.

and thus, the number of durations observed, that are at most length x, was increased significantly.

Further, the empirical durations were closer to the Diggle approximation in terms of the predicted number of shorter durations then in the EDF plot using a theoretical distribution derived from random simulation. The clustering of the empirical events indicated by Figures 5.1.2 and 5.1.3 and the over-estimation of shorter durations by Diggle's (1983) approximation account for this result. Given that the EDF plot based on random simulations **and** the more conservative test of clustering provided by the Diggle (1983) approximation both indicate a clustered pattern of events, we accept the notion that the event pattern is, on the whole, clustered. In the regressive analysis that follows, this information is useful in terms of interpreting the parameter estimates of explanatory variables and their affect on duration length. The results of this graphical analysis also demonstrate that **both** the censored and uncensored data sets show clustering and, at this point, no distinction based on the manner in which durations are measured is reflected in the assessment of the event pattern. With this knowledge of the clustering of Harvestore adopters in space, the analysis progresses into an event history investigation of duration dependency and the role of explanatory variables.

5.2 Event History Modelling

This stage of the analysis considers the distributional form of duration dependence using an event history modelling approach. Initially, description of duration dependence was confined to evaluation the alternative distributional forms of the hazard rate without considering the influence of explanatory variables. It should be noted that inclusion of explanatory variables does not alter the distributional form of duration dependence, but may provide further insights into the most appropriate parametric form to describe the data. Here, the analysis considers the appropriateness of the parametric specifications of the exponential, Weibull, Gompertz and log-logistic distributions in representing the observed durations associated with all 528 events, for the censored and uncensored data sets. The relative fit of these different event history models to the observed censored and uncensored durations was assessed in terms of the log-likelihood estimates.

Briefly, the exponential model is a one parameter distribution which assumes that the hazard rate is independent of duration length. The Weibull is a two parameter model which specifies the dependence of the hazard rate on duration length, and simplifies to the exponential model as a special case. The Gompertz model is also a two parameter model, but, unlike the Weibull, does not reduce to exponential model. The exponential, Weibull and Gompertz models are all monotonic functions whereas the log-logistic model allows for the hazard rate to change direction with duration.

The four parametric models were used to estimate the hazard rates for the duration data using the statistics package LIMDEP. Custom estimation routines existed and were used to evaluate all model types, except the Gompertz. In the case of the

Gompertz model, LIMDEP's estimation routine was very sensitive to starting values using the Davidon, Fletcher and Powell (DFP) algorithm and would rarely converge. To overcome this problem, the log-likelihood equation for the Gompertz model was programmed directly into LIMDEP and the log-likelihood function calculated using the MINIMIZE utility in LIMDEP, and the Berndt, Hall, Hall and Hausman (BHHH) algorithm. However, the custom survival model routines within LIMDEP were based on using the log of duration as the dependent variable. The procedure outlined above for the Gompertz model estimation, however, used duration directly. This had the effect of producing log-likelihood values on a different scales then those of the custom routines. Therefore, to achieve consistency and compare the log-likelihoods, the final models were re-calculated using the MINIMIZE utility.

The log-likelihood values given in Table 5.2.1 indicate that a Weibull model with constant=2.088 and $\alpha=1.460$ provides the best fit to the censored durations, whilst a Weibull model with constant=2.109 and $\alpha=1.422$ provides the best fit to the uncensored durations. Conversely, the exponential model is seen to give the worst fit to the durations in both the censored and uncensored cases which indicates that the constant hazard rate assumption is inappropriate for this data. This also implies that duration dependence does exist. In other words, the findings of our event history analysis thus far is in line with the earlier graphical analysis, that is, the event pattern is not CSR.

To examine duration dependency further, Figure 5.2.1 illustrates the form of the hazard rates for the parametric distributions for censored data (the uncensored data shows very similar hazard rate curves). All of the hazard rates, except, of course, the

Variable	Model Parameter Estimates			
Censored Data	Exponential	Weibull	Gompertz	Log-logistic
constant	2.008 (0.061)	2.088 (0.032)	2.418 (0.071)	1.759 (0.038)
$\alpha/[\sigma]$	1.000	1.460	[0.068]	2.07
Log-likelihood	-1483.4	-1435.0	-1452.6	-1455.7
Uncensored Data				
constant	2.011 (0.060)	2.109 (0.033)	2.369 (0.072)	1.772 (0.038)
$\alpha/[\sigma]$	1.000	1.422	[0.055]	2.082
Log-likelihood	-1589.9	-1543.6	-1562.8	-1566.0

Table 5.2.1: Model parameter estimates with no explanatory variables.

exponential, increase with increasing duration (i.e., they exhibit positive duration dependence). It is important to note that the vast majority of the observations are concentrated in the medium to short durations (see Figure 5.2.2), and so particular attention should be paid to the form of the hazard rates in that range. Further, Figure 5.2.1 shows that, in this range, the Weibull hazard rate lies just below that of the log-logistic, but above the Gompertz. The log-logistic hazard illustrates a rapid rise with duration in this range, then declines slowly with increasing duration. The Gompertz hazard rate increases at an increasing rate with duration, but has a smaller hazard rate then even the exponential in the lower duration range. The positive duration dependence across the models, except the exponential, is consistent with a clustered event pattern of adopters indicated earlier by the graphical analysis. This clustering is best represented by the Weibull hazard rate, primarily because it provides a better fit at the medium to small data range.

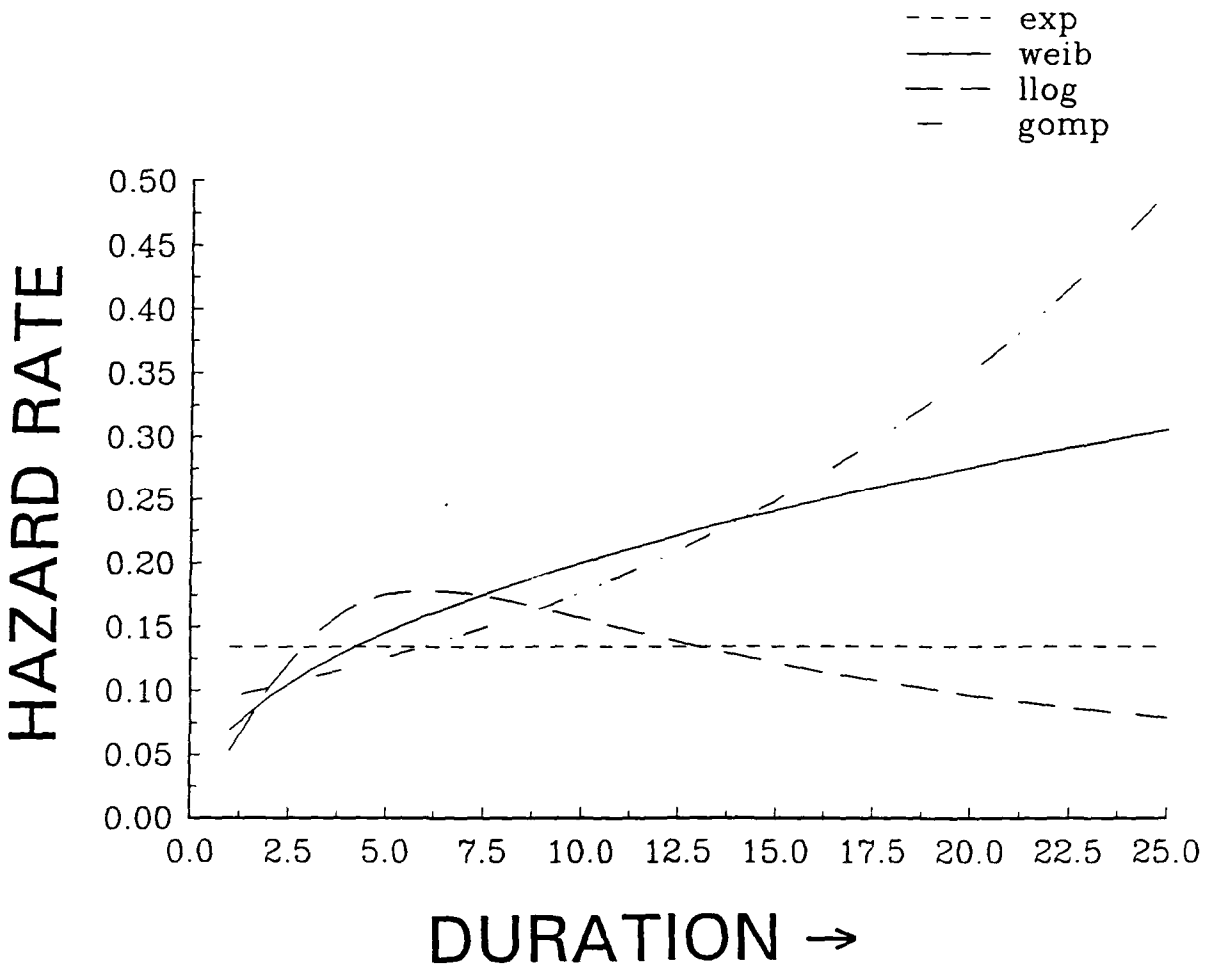

Figure 5.2.1: Hazard rate curves for censored durations.

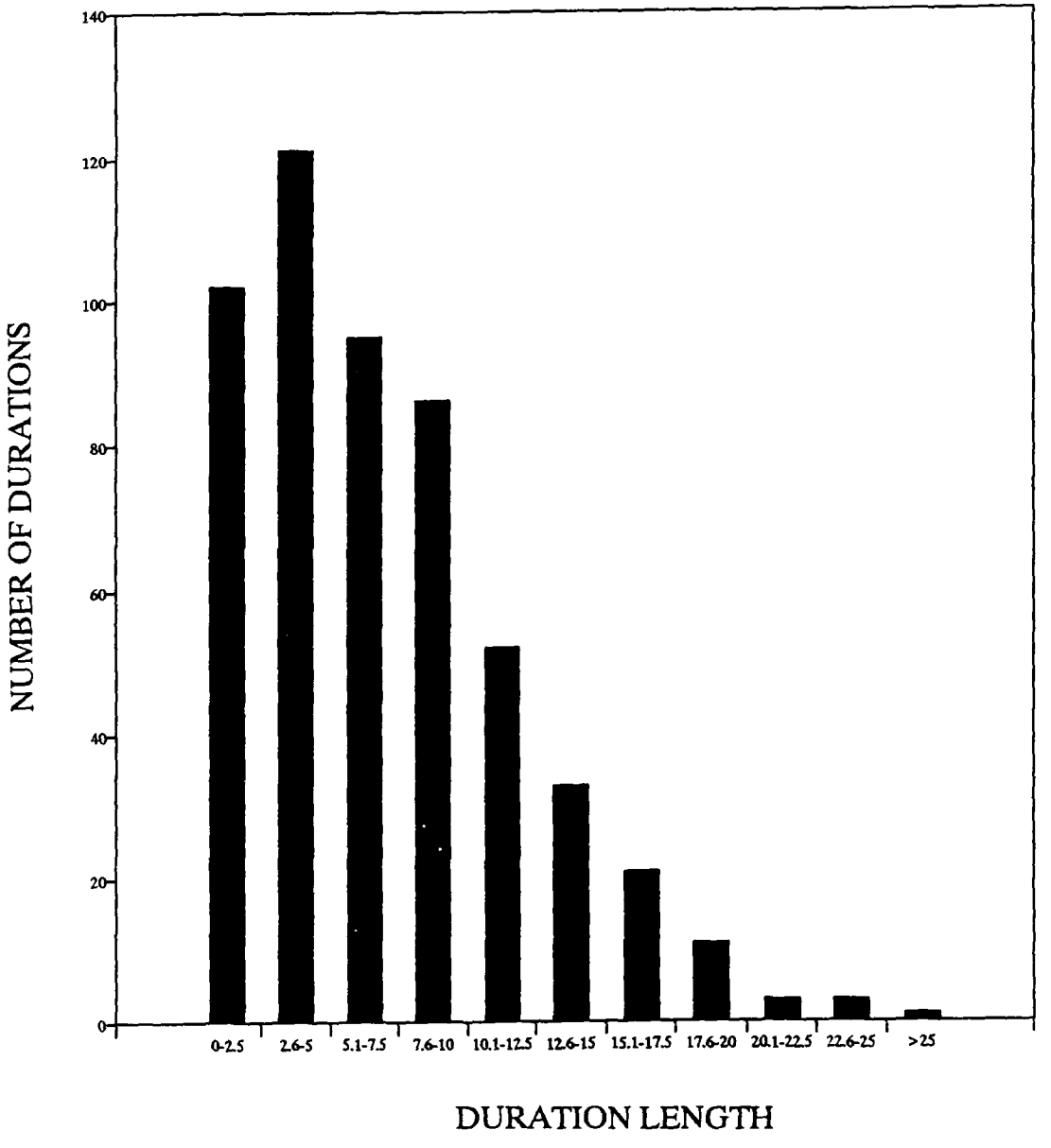

Figure 5.2.2: Histogram of the number of durations at different duration lengths.

5.2.1 Event History Regression Modelling

At this point in the analysis, consideration of the influence of explanatory variables on duration was the focus of attention. Examination of the effects of explanatory variables on duration was initially concerned with the exploration of various combinations of explanatory variables for both the parametric models and the partially parametric Cox model. Recall that the explanatory variables can be examined by including them in the parametric forms previously investigated or alternatively, they can be examined using Cox's (1972) proportional hazards model in which there is no parametric specification of the baseline hazard rate. This exploratory regressive analysis used only the LIMDEP custom routines, hence the Gompertz model was excluded, and attempted to determine which explanatory variables exhibited a significant influence on duration.

Table 5.2.2 shows the parameter estimates of the exponential, Weibull, log-logistic and Cox models for the SAL, DEN, HAY, and COR variables, and for both the censored and uncensored data. For the censored data, the SAL (i.e., the sales agent) variable was significant in all models (except, of course, the exponential) at the 90% level. It is important to note here, that the SAL variable also came out as significant in the Cox model, which reinforces the notion that this explanatory variable is important, since the Cox model makes no assumptions regarding the form of the baseline hazard. The DEN variable, measuring the farming density of the townships within the study area, is significant in the log-logistic model for the censored data, but subsequent parameter estimates for DEN were consistently not significant for all the other models and variable

Variable	Model Parameter Estimates†			
Censored Data	Exponential	Weibull	Log-logistic	Cox*
constant	1.094 (0.962)	1.975 (0.158)	1.400 (0.176)	-
SAL	0.198 (0.124)	0.115‡ (0.063)	0.143‡ (0.074)	0.183‡ (0.095)
DEN	0.079 (0.092)	0.019 (0.047)	0.088‡ (0.051)	0.029 (0.061)
HAY	0.009 (0.334)	-0.145 (0.151)	-0.046 (0.157)	-0.189 (0.182)
COR	-0.112 (0.126)	-0.054 (0.064)	-0.089 (0.076)	-0.088 (0.096)
α	1.000	0.680	0.477	-
Log-likelihood	-669.9	-618.3	-637.8	-2642.2
Uncensored Data				
constant	1.975 (0.298)	2.158 (0.157)	1.472 (0.174)	-
SAL	0.051 (0.124)	0.036 (0.062)	0.100 (0.073)	0.061 (0.091)
DEN	0.006 (0.092)	-0.016 (0.048)	0.077 (0.510)	-0.016 (0.060)
HAY	-0.151 (0.304)	-0.182 (0.159)	-0.062 (0.157)	-0.237 (0.181)
COR	-0.021 (0.124)	-0.009 (0.062)	-0.062 (0.074)	-0.024 (0.093)
α	1.000	0.700	0.477	-
Log-likelihood	-683.5	-636.4	-657.6	-2794.5

* based on Partial Likelihood estimation
† estimated with LIMDEP custom models
‡ significant at the 90% level

Table 5.2.2: Model parameter estimates with explanatory variables.

combinations, including the Cox model. For the uncensored data, none of the parameter estimates were significant. This contrast in results represents the first time in this analysis that measuring censored durations has made a significant difference in results as compared to the uncensored data.

The estimated log-likelihoods for the parametric models with explanatory variables indicate that the Weibull model with constant=1.975 and α=0.680 provides the best fit for the censored durations, whilst the Weibull model with constant=2.158 and α=0.700 best fits the uncensored durations. The exponential models provide the worst fit to both the censored and uncensored durations and this result is not at all surprising given the clustered pattern (i.e., positive duration dependence) of events. The explanatory variables shown in Table 5.2.2 were the most consistent of all the variables in terms of sign, despite only having SAL and DEN as significant. For instance, the HAY (i.e., hay feed type) and COR (i.e., corn feed type) variables exhibited a negative influence on duration and, therefore, increased the hazard rate. The DEN variable, as mentioned above, had an unexpected positive influence on duration for the censored durations, but changed signs in the uncensored case. The remaining farm type variables (i.e., HOG, DAI) and the spatial structure variable, Corn Heat Units, (CHU) were similarly inconsistent in terms of sign, and never statistically significant in terms of their t-ratio, for both the censored and uncensored cases.

At this stage of analysis, the MINIMIZE routines for all the parametric models and the Cox model were used to re-calculate the log-likelihoods, for comparison purposes, including only the SAL variable. Table 5.2.3 shows that the SAL variable is

Variable	Model Parameter Estimates				
Censored Data	Exponential	Weibull	Gompertz	Log-logistic	Cox*
constant	1.935 (0.096)	2.009 (0.049)	2.310 (0.085)	1.673 (0.058)	-
SAL§	0.114 (0.124)	0.123† (0.062)	0.185† (0.092)	0.138‡ (0.074)	0.190† (0.093)
$\alpha/[\sigma]$	1.000	0.681	[0.070]	2.078	-
Log-likelihood	-1482.7	-1433.2	-1450.7	-1454.1	-2643.1
Uncensored Data					
constant	1.973 (0.089)	2.081 (0.047)	2.340 (0.090)	1.708 (0.057)	-
SAL§	0.116 (0.123)	0.045 (0.059)	0.045 (0.084)	0.099 (0.072)	0.073 (0.089)
$\alpha/[\sigma]$	1.000	1.422	[0.055]	2.087	-
Log-likelihood	-1589.9	-1543.4	-1562.7	-1565.2	-2795.4

* based on Partial Likelihood estimation
† significant at the 95% level
‡ significant at the 90% level
§ Sales Agent defined in binary terms: 1=exceptional, 0=typical.

Table 5.2.3: Model parameter estimates including the sales agent variable.

significant at the 95% level for the Cox, Weibull and Gompertz models, and at the 90% level for the log-logistic model, but again, only for the censored data. Not surprisingly, the Weibull provides the best fit to both the censored and uncensored durations. The Gompertz model, which was absent from the previous regressive analysis, is shown to have the next best fit to the Weibull, which is expected given that the Gompertz hazard can be specified as monotonically increasing similar to the Weibull hazard. Given these results, two areas of further investigation were apparent. First, the role of the sales agent variable, and secondly, the discrepancy between the results of the censored and uncensored duration results.

With regards to the role of the sales agent variable, remember that the variable was dichotomized so that SAL=1 referred to sales by *exceptional* sales agents, while SAL=0 indicated Harvestore sales by *typical* sales agents. Given the significant positive parameter estimate of the SAL variable, it appears that Harvestore adopters who purchased their silos from typical sales agents (i.e., those that sold less than 10% of all the Harvestore's in the study area, individually) are associated with shorter durations relative to the adopters who purchased their silos from one of the exceptional sales agents. The positive parameter estimate not only indicates longer durations associated with exceptional sales agents, but also a lower hazard rate. This effect of the SAL variable on the hazard rate is clearly shown in Figure 5.2.3 where the hazard rate for the best fit model, the Weibull, is plotted at SAL=1 and SAL=0. The hazard rate plot shows that the probability of a duration terminating is higher for the typical sales agents (SAL=0), then the for the exceptional sales agents (SAL=1), since the hazard associated

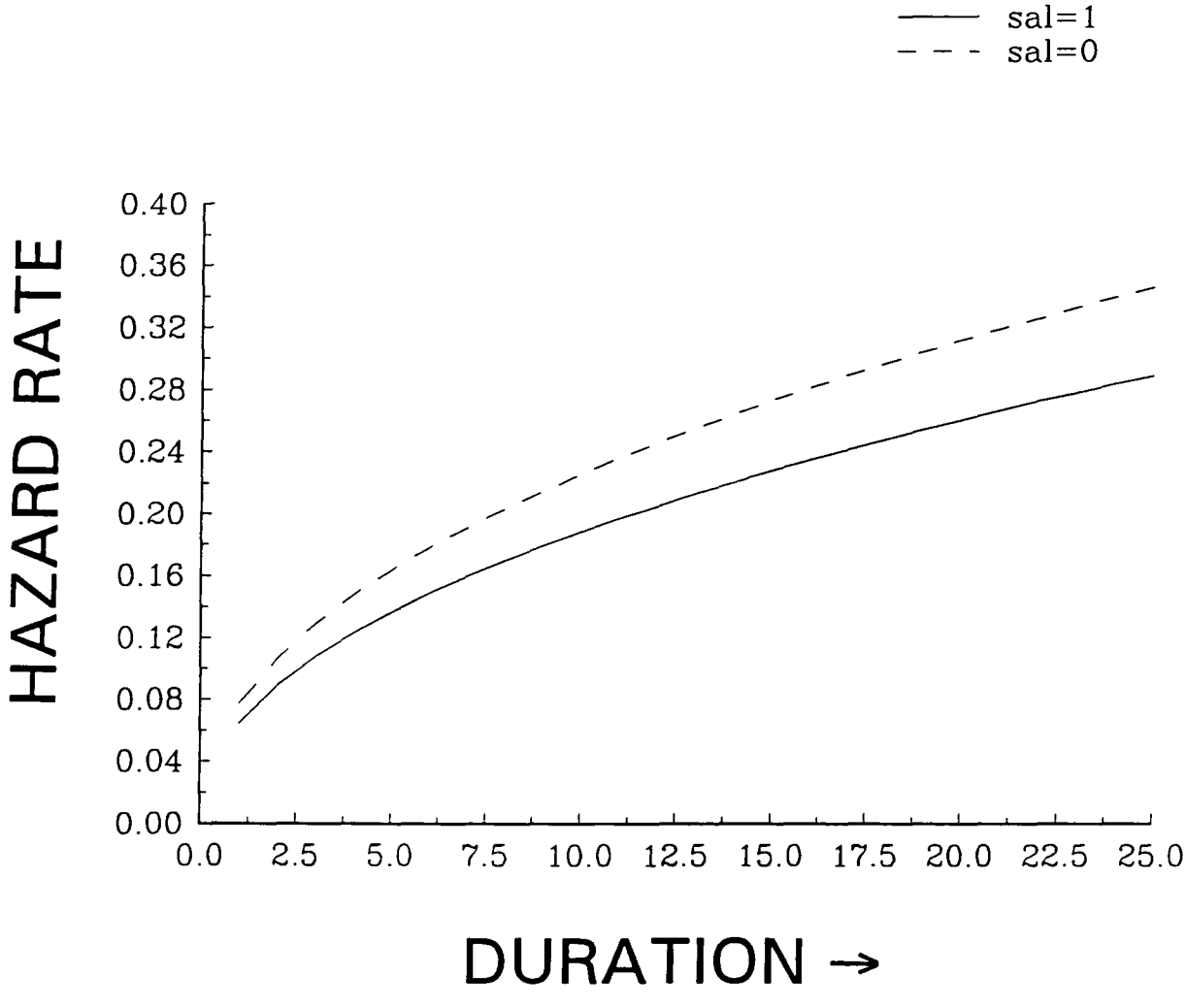

Figure 5.2.3: Hazard rate curves for censored Weibull model including sales agent.

with the former is above that of the exceptional sales agents. This also means that the typical sales agents are associated with shorter nearest neighbour durations as opposed to the exceptional sales agents, who are associated with longer durations.

The implication from the findings for the SAL variable lend support to Scorgie's (1992) hypothesis that the sales agent was one of, if not the, major contagion in the spread of the Harvestore in southern Ontario. The results also indicate that the exceptional and typical sales agents had different strategies and spatial mechanisms of information flow in regards to selling the Harvestore silo. The typical sales agents likely worked in small, well-defined boundaries within the study area, whilst the exceptional sales agents likely had large, dispersed sales territories to account for the number of Harvestore sales they generated, and the longer durations associated with the location of their sales. The typical sales agents probably relied more heavily on the neighbourhood effect of diffusion than the three exceptional sales agents who, taken together, far out-sold all the other sales agents. Of course, it should be noted that some of the typical agents might have only worked for a short time and, therefore, sold very few silos to, most likely, a localized group of farmers.

However, the above result for the SAL variable was only obtained in the censored data case. When each of the events was forced to find a nearest neighbour event within the study area (i.e., the uncensored data), none of the variables, not even the SAL variable, were significant. To further examine this result, some descriptive statistics on the dependent variable (i.e., the durations) for the censored and uncensored cases provided some insight into these conflicting results. Tables 5.2.4 (a) and (b) compare

the durations associated with each event in the censored and uncensored cases. Table 5.2.4 (a) shows the mean duration for the 325 events sold by the exceptional sales agents as being slightly shorter than the mean duration for the typical sales agents (203 events). This is true for the censored and uncensored data, although, as expected, the mean durations in the uncensored data are larger. This result is consistent with the parameter estimates, but does not explain the difference in results between the censored and uncensored data since both data sets show that longer durations are associated with the exceptional sales agents. Table 5.2.4 (b) shows the mean durations again, but only for the 35 boundary events (i.e., those events whose duration would end at the boundary in the censored data). In this case, the censored data mean durations are consistent with the general findings, but the uncensored data mean durations for these boundary events indicate the exact opposite trend. The means for SAL=1 and SAL=0 are higher, as expected when durations are forced to find neighbours in the study area, but now the mean duration for the typical sales agents (i.e., 17.233) is much higher relative to the mean duration for the exceptional sales agents (i.e., 13.709).

The dramatic difference in the mean durations for the 35 boundary events clearly accounts for the differing regression results between the censored and uncensored data sets. However, the question remains as to which result is revealing more of the true story behind the process that lead to the event pattern of Harvestore adopters. Certainly, by incorporating the information on censored durations in the estimation of model parameters and attaining two very different results, one showing a significant effect of the sales agent on duration and the other no significant effect, we are forced to

(a)

SAL=1; n=325	Censored	Uncensored
mean std. dev.	7.386 5.109	7.647 5.364
SAL=0; n=203		
mean std. dev.	6.280 4.478	7.192 5.629

(b)

SAL=1; n=19	Censored	Uncensored
mean std. dev.	8.409 4.748	13.709 6.334
SAL=0; n=16		
mean std. dev.	7.495 6.951	17.233 7.955

Table 5.2.4 (a) Mean durations for all events, (b) mean durations for boundary events.

reconsider simply ignoring censored observations in this type of regressive analysis. In fact, these results provide a strong argument in favour of some sort of sensitivity analysis between censored and uncensored data sets.

As with any regressive analysis, the findings here are subject to some limitations. For instance, it is likely that the effects of farm and feed type dummy variables were either dampened or emphasized by the likelihood of spatial heterogeneity in the distribution of the various farm and feed types in the study area. This notion is emphasized by the lack of statistical significance of either spatial structure variable, the farm operator density (DEN) and the spatial variation in the level of Corn Heat Units (CHU), indicating that other variable measuring variation in spatial structure might have been more appropriate. For example, the distance from each event to the nearest urban centre may have better accounted for the farm density around each event. Thus, the consistent, although not statistically significant, negative influence of the Corn and Hay feed types might be either the result of spurious heterogeneity in the study area, given that the spatial structure variables were not significant, or might indicate stronger communication ties, and therefore, more adoptions amongst farmers using corn and hay feed.

5.3 Thesis Conclusions

This thesis sought to extend the research of Odland and Ellis (1992) in applying event history models to the analysis of spatial point patterns. Although methodological in nature, an empirical investigation of the event pattern resulting from the spatial diffusion of an agricultural innovation, the Harvestore, was an important component of this research. The thesis began with a review of some distance based methods of spatial point pattern analysis which are analogous, in terms of their use of frequency distributions of intervals (or durations) between events, to event history methods. Chapter Two revealed some of the problems associated with distance based methods of spatial point pattern analysis, notably nearest neighbour statistics. Of particular interest in this thesis was the problem of edge effects, which tends to increase the mean nearest neighbour distance between events. Other issues of concern from this review of literature included the notion that nearest neighbour durations, although intuitively appealing for measuring spatial dependencies between events, are only one subset within a multitude of possible inter-event durations that may be analyzed to investigate a spatial event pattern. Also, it was noted that distance based methods did not provide a regressive component to investigate the influence of explanatory variables on duration lengths.

The issues of defining an appropriate duration to model, and potential explanatory variables related to these durations was the focus of section 2.3 on the spatial diffusion of agricultural innovations. In this regard, the empirical focus of this thesis was used to guide the specification of an appropriate duration definition and explanatory variables.

After considering modelling durations based on the temporal sequence of events (i.e., adopters), traditional nearest neighbour distances were chosen as the most appropriate duration to model. This decision was based on the emphasis in this thesis on the spatial mechanisms of the diffusion process, where events close in spatial proximity are though to be more related, in terms of information flow regarding the innovation, then events further apart. In addition, several explanatory variables were derived from the literature on the spatial diffusion of agricultural innovations, including the spatial mechanisms of interpersonal contact.

In Chapter Three, the methods of event history data analysis that were used for the analysis of the event pattern of adopters were introduced. Here, the spatial durations could be viewed in as cohort survival data in temporal analysis terms. In other words, the durations are one-dimensional and have a common start point. This chapter also formally introduced spatial censoring as a means of accounting for some edge effects. In this regard, spatial censoring is analogous to "right-censoring" in temporal terms and may be defined as the shortest distance from a boundary event to the boundary, and thus, it does not terminate at a nearest neighbour event. Recall that boundary events were defined as those events closer to the boundary than any internal event location. The concepts of right censoring and spatial censoring are similar in that they both convey the notion that the censored observation in not fully known, but the analyst has certain knowledge that the duration extends to at least the observed length. Spatially censored observations can be incorporated into model estimation, and, therefore, can be readily compared to a data set consisting entirely of uncensored durations.

The primary interest of this research is concerned with evaluating the distributional form of duration dependence in the event pattern. Duration dependency corresponds to one type of spatial state dependence, and in this thesis, was defined as the length of the nearest neighbour duration between adopter farms. That is, the probability of an event occurring is modelled as dependent on the length of the nearest neighbour duration. The results of the graphical analysis indicated that the event pattern was clustered and so positive duration dependency in the data was expected. In the context of the event history models used in this thesis, the distributional form of this duration dependence can be expressed in terms of the hazard rate.

The main findings of this thesis relate to the performance of alternative event history models in representing the duration dependence in the event pattern. Results indicate that a hazard rate derived from an monotonically increasing Weibull model best described both the censored and uncensored data. The exponential model, as expected, consistently gave the worst fit to both sets of durations, which, of course, indicated that a constant hazard rate assumption was inappropriate for this event pattern. The other main finding in this thesis concerns the regressive analysis. Several interesting results at this stage of the analysis illustrate both the important contribution event history methods provide to spatial point pattern analysis, and the significance of the research in this thesis. For instance, the sales agent variable was shown to have a significant effect on duration, illustrating their role as the contagion in the diffusion of the innovation. This result was reinforced by the fact that identical results were derived from the Cox model, which made no assumption about the baseline hazard. Further, the *exceptional*

sales agents were found to have longer durations associated with them (i.e., a positive effect on duration) and therefore, decreased the hazard rate. This result implied that the *typical* sales agents had more localized selling patterns and likely used the neighbourhood effect to influence potential adopters. Conversely, the exceptional sales agents had more dispersed sales territories and likely relied less on the neighbourhood effect for selling the Harvestore to farmers.

Another important finding in regards to the regressive analysis was the fact that the sales agent variable was only significant for the censored data. The fact that the coefficient estimate was significant in the censored case, and not significant in the uncensored case, illustrates the sensitivity of model parameter estimates to edge effects. An evaluation of the mean durations associated with the boundary events in the censored and uncensored data revealed that the uncensored durations (which were forced to find neighbours in the study area) reversed the trend found throughout the rest of the durations in that the durations associated with the typical sales agents now were longer than those associated with the exceptional sales agents. This finding suggests that some sensitivity analysis be undertaken, in terms of measuring both censored and uncensored durations, in other regressive analyses of this nature.

5.4 Directions for Future Research

In many ways, research directed at methodological issues related to the adaptation of methods from one discipline or one domain to another, as in the case of this research, and other work by geographers (e.g., spatial interaction modelling), the aim is to raise

as many questions as answers and uncover some of the salient issues that must be examined on the road to incorporating that methodology into the arsenal of tools presently available. In bringing event history models into the spatial domain, Odland and Ellis (1992) and, hopefully, the research reported here, have set a clear research agenda for geographers and other spatial scientists.

Many issues for further investigation are evident from this thesis. The most obvious involves evaluating the effectiveness of incorporating spatially censored observations. The method employed in this thesis is intuitively appealing because it uses certain knowledge about the minimum nearest neighbour distance from an event, and this observation is incorporated in model estimation. However, an interesting assessment of the usefulness of censored observations could be examined through an analysis similar to the one undertaken in this thesis, but with the addition of having actual information on events outside of the boundary, using, perhaps, some sort of buffer zone of events around the study area. In this way, the censored, uncensored and actual durations may be analyzed, and the results compared to see if the censored durations provide a better depiction of the actual durations then the uncensored (or traditional nearest neighbour) durations. Given that methods of model estimation in the presence of censored data already exist, a thorough evaluation of their usefulness is important.

Research issues incorporating both temporal and spatial event history modelling are another area of consideration for further work. In this vein, we can envision a data set, not unlike the one used in this thesis, containing both the exact timing and location of events, as well as some spatial and temporal explanatory variables. Temporal hazard

rates can then be estimated separately from spatial hazard rate estimates, and this information could eventually be brought together to provide probability surface contours. Thus, for selected discrete time intervals, a map of event probabilities could be constructed illustrating areas most or least likely to experience an event. In the case of some natural disasters and contagious diseases, where the location of an event will effect the timing and location of the next event, such an analysis is significant. Certainly, the advent of Temporal GIS, where temporal changes to geographic phenomena are stored and visually represented with hypermedia (i.e., sound and changing colours) may well be an avenue to operationalize such methods.

Finally, it is important to note that the event history models used in this thesis represented only a small subset of the available methods of longitudinal analysis. On the basis of the interesting findings in this thesis and those by Odland and Ellis (1992), it seems obvious that entirely new avenues can be pursued in terms of the linkage between spatial point pattern analysis and event history modelling. Odland and Ellis (1992) looked at spatially varying explanatory variables and spatial heterogeneity, and this thesis considered duration dependency with a range of parametric models, spatial censoring and the effects of a series of explanatory variables (using parametric and partially parametric models), but, taken together, these efforts merely scratch the surface of techniques available in the expanding literature on event history modelling.

REFERENCES

Allison, P.D. (1982) Discrete-Time Methods for the Analysis of Event Histories in Sociological Methodology Ed. S. Leinhardt (Jossey-Bass, San Francisco, CA) pp.61-97.

------- (1984) Event History Analysis: Regression for Longitudinal Event Data. (Sage, Beverly Hills).

Agterberg, F.P. (1984) Trend Surface Analysis in Spatial Statistics and Models Eds. G.L. Gaile and C.J. Willmott (Dordrecht, Reidel) pp.147-171.

Aronoff, S. (1989) Geographic Information Systems: A Management Perspective. (WDL Publications, Ottawa).

Bartlett, M.S. (1963) "The Spectral Analysis of Point Processes", Journal of the Royal Statistical Society B 25, pp.264-296.

------- (1964) "The Spectral Analysis of Two-Dimensional Point Processes", Biometrika 51, pp.299-311.

Batty, M.J. (1976) Urban Modelling: Algorithms, Calibrations, Predictions. (Cambridge University Press, Cambridge).

Besag, J.E, (1978) "Some Methods of Statistical Analysis for Spatial Data", Bulletin of the International Statistical Institute, 47, pp.2, 77-92.

Blossfeld, H.P., Hamerle, A. and Mayer, K.U. (1989) Event History Analysis. (Lawrence Erlbaum Associates, New Jersey).

Boots, B. (1981) "Weighting Theissen Polygons", Economic Geographer 57, pp.248-259.

------- and Murdoch, D.J. (1983) "The Spatial Arrangement of Random Voronoi Polygons", Computers and Geosciences 9, pp.351-365.

------- and Getis, A. (1988) Point Pattern Analysis. (Sage Publications, Newbury Park).

Brown, L.A. (1981) Innovation Diffusion: A New Perspective. (Methuen, New York).

Burrough, P.A. (1986) Principles of Geographical Information Systems for Land Resources Assessment. (Clarendon Press, Oxford).

Cancian, F. (1979) The Innovator's Situation: Upper-Middle-Class Conservatism in Agricultural Communities. (Stanford University Press, Stanford, California).

Carlstein, T. (1978) Innovation, Time Allocation and Time-Space Packing in Human Activity and Time Geography Eds. T. Carlstein, D. Parkes, and N. Thrift (Edward Arnold, London).

Chorley, R.J. and Haggett, P. (1965) "Trend Surface Mapping in Geographical Research", Transactions of the Institute of British Geographers 37, pp.47-67.

Clark, P.J. and Evans, F.C. (1954) "Distance to Nearest Neighbour as a Measure of Spatial Relationships in Populations", Ecology 35, pp.445-453.

Clark, W.A.V. (1992) "Comparing Cross-sectional and Longitudinal analyses of Residential Mobility and Migration", Environment and Planning A 24, pp.1291-1302.

Cliff, A.D. and Ord, J.K. (1981) Spatial Processes: Models and Applications. (Pion, London).

-------, Haggett, P., Ord, J.K., and Versey, G.R. (1981) Spatial Diffusion: An Historical Geography of Epidemcs in an Island Community. (Cambridge University Press, Cambridge).

Colman, G.P. (1968) "Innovation and Diffusion in Agriculture", Agricultural History 42, pp.173-187.

Cox, D.R. (1972) "Regression Models and Life-Tables (with discussion)", Journal of the Royal Statistical Society B 24, pp.406-424.

------- (1975) "Partial Likelihood", Biometrika 62, pp.269-276.

------- and Oakes, D. (1984) Analysis of Survival Data. (Chapman and Hall, Andover).

Crain, I.K. (1972) "Monte Carlo Simulation of Random Voronoi Polygons: Preliminary Results", Search 3, pp.220-221.

------- (1978) "The Monte-Carlo Generation of Random Polygons", Computers and Geosciences 4, pp.131-141.

Crouchley, R.(Ed.) (1987) Longitudinal Data Analysis. (Sage, Beverly Hills, CA).

-------, Pickles, A.R. and Davies, R.B. (1982a) "Dynamic Models of Shopping Behaviour: Testing the Linear Learning Model and Some Alternatives", Geografiska Annaler B pp.27-33.

Dacey, M.F. (1975) "Evaluation of the Poisson Approximation to Measures of the Random Pattern in a Square", Geographical Analysis 7, pp.351-367.

Davies, R.B. (1984) "A Generalized Beta-Logistic Model for Longitudinal Data with an Application to Residential Mobility", Environment and Planning A 16, pp.1375-86.

------- (1987) Mass Point Methods for Dealing with Nuisance Parameters in Longitudinal Studies in Longitudinal Data Analysis Ed. R. Crouchley (Gower, Aldershot).

------- (1988) A Reappraisal of Some Simple Statistical Models in London Papers in Regional Science 18. Longitudinal Data Analysis: Methods and Applications Ed. M. Uncles (Pion, London) pp.103-115.

------- and Crouchley, R. (1984) "Calibrating Longitudinal Models of Residential Mobility and Migration", Regional Science and Urban Economics 14, pp.231-247.

------- and ------- (1985) "Control for Omitted Variables in the Analysis of Panel and Other Longitudinal Data", Geographical Analysis 17, pp.1-15.

-------, ------- and Pickles, A.R. (1982a) "A Family of Hypotheses Tests for a Collection of Short Event Series with an Application to Female Employment Participation", Environment and Planning A 14, pp.603-614.

-------, -------, and -------. (1982b) "Modelling the Evolution of Heterogeneity in Residential Mobility", Demography 19, pp.291-299.

------- and Pickles, A.R. (1983) "The Estimation of Duration-of-Residence Effects: A Stochastic Modelling Approach", Geographical Analysis 15, pp.305-317.

------- and ------- (1985) "Longitudinal Versus Cross-sectional Methods for Behavioural Research: a First-Round Knockout", Environment and Planning A 17, pp.1315-1329.

------- and ------- (1987) "A Joint Trip Timing Store-Type Choice Model for Grocery Shopping, Including Inventory Effects and Nonparametric Control for Omitted Variables", Transportation Research A 21, pp.345-361.

------- and ------- (1991) "An Analysis of Housing Careers in Cardiff", Environment and Planning A 23, pp.629-650.

-------, ------- and Crouchley, R. (1982a) "Some Methods for the Testing and Estimation of Dynamic Models Using Panel Data", Environment and Planning A 15, pp.1475-1488.

-------, -------, and -------. (1982b) "Event History Testing: Effects in a Collection of Event Series", Sociological Methods and Research 10(3), pp.285-302.

Dawson, A.H. (1975) "Are Geographers Indulging in a Landscape Lottery?", Area 7, pp.42-45.

DeTemple, D.J. (1970) A Space Preference Approach to the Determination of Individual Contact Fields in the Spatial Diffusion of Harvestore Systems in Northeast Iowa. Department of Geography, Michigan State University, PH.D. dissertation.

------- (1971) A Space Preference Approach to the Diffusion of Innovations: The Spread of Harvestore Systems Through Northeast Iowa. (Geographic Monograph Series Department of Geography, Indiana University).

Diggle, P. (1979) Statistical Methods for Spatial Point Patterns in Ecology in Spatial and Temporal and Analysis in Ecology Eds. R.M. Cormack and J.K. Ord (Fairland, International Co-Operative Publishing House) pp.95-150.

------- (1983) Statistical Analysis of Spatial Point Patterns. (Academic Press, New York).

Eastman, R.J. (1988) "Idrisi: A Geographic Analysis System for Research Applications", The Operational Geographer 15, pp.18-21.

------- (1991) Idrisi: A Grid-Based Geographic Analysis System Version 3.2. (Clark University Graduate School of Geography, Massachusetts).

Flinn, C.J. and Heckman, J.J. (1982) New Methods for Analyzing Individual Event Histories in Sociological Methodology 1982 Ed. S. Leinhardt (Jossey-Bass, San Francisco, CA) pp.99-140.

Fotheringham, A.S. (1993) "GIS and Exploratory Spatial Data Analysis (forthcoming).

Fotheringham, A.S. and Rogerson, P. (1993) "GIS and Spatial Analytical Problems", International Journal of Geographical Information Systems. (forthcoming).

Getis, A. and Boots, B. (1978) Models of Spatial Processes. (Cambridge University Press, Cambridge).

------- (1983) "Second-Order Analysis of Point Patterns: The Case of Chicago as a Multi-Center Urban Region", Professional Geographer, 35, pp.73-80.

Gould, P.R. (1969) Spatial Diffusion. (Association of American Geographers, Resource Paper Series, Washington).

Greene, W.H. (1990) LIMDEP: Version 5.1, Econometric Software Inc., N.Y.

Greig-Smith, P. (1964) Quantitative Plant Ecology 2nd Edition. (Butterworth, London).

Griffith, D. and Amrhein, C. (1983) An Evaluation of Correction Techniques for Boundary Effects in Spatial Statistical Analysis: Traditional Methods Geographical Analysis 15, pp.352-360.

Gross, N. (1949) "The Differential Characteristics of Acceptors and Non-Acceptors on an Approved Agricultural Practice", Rural Sociology 14, pp.148-156.

Haining, R. (1990) Spatial Data Analysis in the Social and Environmental Sciences. (Cambridge University Press, Cambridge).

Hägerstrand, T. (1952) The Propagation of Innovation Waves. (Lund, Gleerup, Lund Studies in Geography).

------- (1965a) "A Monte Carlo Approach to Diffusion", Archives Europeennes de Sociologie, 6, pp.43-67.

------- (1965b) Quantitative Techniques for Analysis of the Spread of Information and Technology in Education and Economic Development Eds. C.A. Anderson, and M.J. Bowman (Chicago, Aldine).

------- (1967a) Innovation Diffusion as a Spatial Process. (Chicago, University of Chicago Press).

------- (1967b) On the Monte Carlo Simulation of Diffusion in Quantitative Geography, Part 1: Economic and Cultural Topics Eds. W.L. Garrison, and D.F. Marble (Evanston, Northwestern University Press, Studies in Geography).

Halperin, W.C. (1985) The Analysis of Panel Data for Discrete Choices in Measuring the Unmeasurable Eds. P. Nijkamp, H. Leitner, and N.Wrigley (Martinus Nijhoff Dordrecht) pp.561-85.

Hannan, M.T. (1984) Multi-State Demography and Event-History Analysis in Stochastic Modelling of Social Processes Eds. A. Diekmann and P. Mitter (Academic Press, Orlando) pp.39-88.

------- and Tuma, N.B. (1979) "Methods for Temporal Analysis", Annual Review of Sociology 5, pp.303-328.

Harvey, D.W. (1968) "Pattern Process and the Scale Problem in Geographical Research", Transactions and Papers, Institute of British Geographers, 45 pp.71-78.

------- (1969) Explanation in Geography. (Edward Arnold, London).

Havens, A.E. (1965) "Increasing the Effectiveness of Predicting Innovativeness", Rural Sociology 30, pp.150-165.

Heckman, J.J. (1981a) Statistical Models for Discrete Panel Data in Structural Analysis of Discrete Data with Econometric Applications Eds. C. Manski, D. McFadden (MIT Press, Cambridge, MA) pp.114-178.

------- (1981b) The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Stochastic Process and Some Monte Carlo Evidence on their Practical Importance in Structural Analysis of Discrete Data with Econometric Applications Eds. C. Manski, D. McFadden (MIT Press, Cambridge, MA) pp.179-195.

------- and Borjas, G. (1980) "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence", Economica 47, pp.247-283.

------- and Singer, B. (1980) Population Heterogeneity in Demographic Models in Multidimensional Mathematical Demography Eds. K.C. Land and A. Rogers (Academic Press, New York) pp.567-599.

------- and ------- (1982) The Identification Problem in Econometric Models for Duration Data in Advances in Econometrics Ed. W. Hilderbrand (Cambridge University Press, Cambridge) pp.39-77.

------- and ------- (1984a) "A Method for Minimising the Impact of Distributional Assumptions in Econometric Models for Duration Data", Econometrica 52, pp.271-320.

------- and ------- (1984b) "The Identifiability of the Proportional Hazard Model", Review of Economic Studies 60(2), pp.231-243.

------- and ------- (1984c) "Econometric Duration Analysis", Journal of Econometrics 24, pp.63-132.

------- and ------- (1986) Econometric Analysis of Longitudinal Data in Handbook of Econometrics Eds. Z. Griliches and M.D. Intriligator (North Holland, Amsterdam) 1689-1763.

Hinde, A.L. and Miles R.E. (1980) "Monte Carlo Estimates of the Distributions of the Random Polygons of the Voronoi Tessellation with respect to a Poisson Process", Journal of Statistical and Computer Simulations 10, pp.205-223.

Hsiao, C. (1986) Analysis of Panel Data. (Cambridge University Press, Cambridge).

Hudson J.C. and Fowler, P.M. (1966) "The Concept of Pattern in Geography." University of Iowa, Department of Geography, Discussion Paper No.1.

Kalbfleisch, J.D. and Prentice, R.L. (1980) The Statistical Analysis of Failure Time Data. (John Wiley, New York).

Katz, E. (1957) "The Two-Step Flow of Communications: An Up-to-Date Report on an Hypothesis", Public Opinion Quarterly 21, pp.61-78.

Kiefer, N.M. (1988) "Economic Duration Data and Hazard Functions", Journal of Economic Literature 26, pp.646-679.

King, L.J. (1962) "A Quantitative Expression of the Pattern of Urban Settlement in Selected Areas of the United States", Tijdschrift voor Economische en Sociale Geografie 53, pp.1-7.

------- (1976) "Alternatives to a Positive Economic Geography", Annals of the Association of American Geographers 66, pp.293-308.

Lawless, J.F. (1982) Statistical Models and Methods for Lifetime Data. (Wiley, New York).

Lazarfeld, P.F. and Merton, R.K. (1964) Friendship as Social Process: A Substantive and Methodological Analysis in Freedom and Control in Modern Society Eds. M. Berger and others (Octagon, New York).

Lionberger, H.F. (1960) Adoption of New Ideas and Practices. (Iowa State University Press, Iowa).

Mahajan, V. and Peterson, R. (1985) Models for Innovation Diffusion. (Sage Publications, Beverly Hills).

Malecki, E.J. (1975) Innovation Diffusion Among Firms. Department of Geography, Ohio State University, Ph.D. dissertation.

Mayer, K.U. and Tuma, N.B. (1990) Event History Analysis in Life Course Research. (University of Wisconsin Press, Wisconsin).

Morrill, R.L. and Pitts, F.R. (1967) "Marriage, Migration and the Mean Information Field", Annals Association of American Geographers 57, pp.401-422.

------- (1974) Growth Center-Hinterland Relations in Proceedings of the Commission on Regional Aspects of Economic Development of the International Geographical Union, Vol.II: Spatial Aspects of the Development Process Eds. F. Helleiner and W. Stohr (Allister, Toronto).

Odland, J. (1988) Spatial Autocorrelation. (Sage, Newbury Park).

------- and Ellis, M. (1992) "Variations in the Spatial Pattern of Settlement Locations: An Analysis based on Proportional Hazards Models", Geographical Analysis 24: 97-109.

Openshaw, S. (1992) "Two exploratory space-time-attribute pattern analysers relevant to GIS" NCGIA Initiative 14: Position Papers.

Openshaw, S. (1991) "Developing appropriate spatial analysis methods for GIS in D.Maguire, M.F. Goodchild and D.W. Rhind (Eds.) Geographical Information Systems: Principles and Applications. (Longman, London) pp. 389-402.

Peuquet, D.J. and Marble, D.F. (Eds.) (1990) Introductory Reading in Geographic Information Systems. (Taylor and Francis, London).

Pickles, A.R. (1983) "The Analysis of Residence Histories and Other Longitudinal Panel Data: a Continuous Time Mixed Markov Renewal Model Incorporating Exogenous Variables", Regional Science and Urban Economics 13, pp.271-285.

-------, Crouchley, R. and Davies, R.B. (1982) "Non-Participants in Choice Processes: An Application to Intra-Urban Migration", Area 14, pp.43-50.

------- and Davies, R.B. (1984) Recent Developments in the Analysis of Movement and Recurrent Choice in Spatial Statistics and Models Eds. G. L. Gaile and C. J. Willmott (Reidel, Netherlands) pp.321-43.

------- and ------- (1985) "The Longitudinal Analysis of Housing Careers", Journal of Regional Science 25, pp.85-101.

-------, ------- and Crouchley, R. (1982) "Heterogeneity, Non-Stationarity and Duration-of-Stay Effects in Migration", Environment and Planning A 14, pp.615-622.

Reader, S. (1988) Incorporating Unobserved Heterogeneity into Longitudinal Models of Repeated Choice: Optimization, Estimation and Simulation Issues. Department of Geography, University of Bristol, unpublished Ph.D. dissertation.

------- and Uncles, M.D. (1988) The Collection and Analysis of Consumer Data in London Papers in Regional Science 18. Longitudinal Data Analysis: Methods and Applications Ed. M. Uncles (Pion, London) pp.45-57.

Ripley, B.D. (1976a) "The Second-Order Analysis of Stationary Point Processes", Journal of Applied Probability 13, pp.255-266.

------- (1977) "Modelling Spatial Patterns (With Discussion)", Journal of the Royal Statistical Society B39, pp.172-212.

------- (1979a) "Simulating Spatial Patterns: Dependent Samples from a Multivariate Density", Applied Statistics 28, pp.109-112.

------- (1981) Spatial Statistics. (New York, Wiley).

Roder, W. (1975) "A Procedure for Assessing Point Patterns Without Reference to Area or Density", The Professional Geographer 27, pp.432-440.

Rogers, E.M. and Shoemaker, F.F. (1971) Communications of Innovations: A Cross Cultural Approach. (Free Press, New York).

------- (1983) Diffusion of Innovations Third Edition. (The Free Press, New York).

Ryan, B. and Gross, N.C. (1943) "The Diffusion of Hybrid Seed Corn in Two Iowa Communities", Rural Sociology 8, pp.15-24.

Scholten, H.J. and Stillwell, J.C.H. (Eds.) (1990) Geographical Information Systems for Urban and Regional Planning. (Kluwer Academic Publishers, Dordrecht).

Scorgie, E.K. (1973) The Diffusion of Harvestore Structures in Southwestern Ontario. Department of Geography, University of Western Ontario, M.A. dissertation.

------- (1992) Adopting Harvestore Systems: A Study of Change, Technology and Diffusion in Agriculture. Department of Geography, University of Waterloo, ongoing Ph.D. Research.

Singer, B. (1981) Estimation of Nonstationary Markov Chains from Panel Data in Sociological Methodology 1981 Ed. S. Leinhardt (Jossey-Bass, San Francisco) pp.319-337.

Star, J. and Estes, J.E. (1990) Geographic Information Systems: An Introduction. (Prentice-Hall, Inc., New Jersey).

Stoyan, D., Kendall, W.S. and Mecke, J. (1987) Stochastic Geometry and its Applications. (Akademie-Verlag, Berlin).

Taylor, P.J. (1977) Quantitative Methods in Geography. (Houghton Mifflin, Boston).

Tobler, W.R. (1969) "Geographical Filters and their Inverses", Geographical Analysis 1, pp.234-253.

Tomlin, C.D. (1990) Geographic Information Systems and Cartographic Modeling. (Prentice-Hall Inc., New Jersey).

Tuma, N.B. (1982) Non-Parametric and Partially Parametric Approaches to Event History Analysis in Sociological Methodology 1982 Ed. S. Leinhardt (Jossey-Bass, San Francisco) pp.1-60.

------- and Hannan, M.T. (1978) Approaches to the Censoring Problem in Analysis of Event Histories in Sociological Methodology 1979 Ed. K.F. Schuessler (Jossey-Bass, San Francisco) pp.209-240.

------- and ------- (1984) Social Dynamics. (Academic Press, Orlando).

Uncles, M.D. (1987) "A Beta-logistic Model of Mode Choice: Goodness of Fit and Inter-temporal Dependence", Transportation Research B 21, pp.195-205.

------- (1988) Issues in Longitudinal Data Analysis in London Papers in Regional Science 18, Longitudinal Data Analysis: Methods and Applications Ed. M.D. Uncles (Pion, London) pp.1-12.

------- (Ed.) (1988b) London Papers in Regional Science 18. Longitudinal Data Analysis: Methods and Applications. (Pion, London).

Unwin, D. (1981) Introductory Spatial Analysis. (Methuen, London).

Upton, G. and Fingleton, B. (1985) Spatial Data Analysis by Example Volume 1: Point Patterns and Quantitative Data. (Wiley, New York).

Vincent, P. (1976) "The General Case: How Not To Measure Spatial Point Patterns", Area 8, pp.161-163.

Wrigley, N. (1985) Categorical Data Analysis for Geographers and Environmental Scientists. (Longman, London).

------- (1986) "Quantitative Methods: the Era of Longitudinal Data Analysis", Progress in Human Geography 10, pp.84-102.

------- and Dunn, R. (1984a) "Stochastic Panel-Data Models of Urban Shopping Behaviour: 1 Purchasing at Individual Stores in a Single City", Environment and Planning A 16, pp.629-50.

------- and ------- (1984b) "Stochastic Panel-Data Models of Urban Shopping Behaviour: 2 Multistore Purchasing Patterns and the Dirichlet Model", Environment and Planning A 16, pp.759-78.

------- and ------- (1984c) "Stochastic Panel-Data Models of Urban Shopping Behaviour: 3 The Interaction of Store Choice and Brand Choice", Environment and Planning A 16, pp.1221-36.

------- and ------- (1985) "Stochastic Panel-Data Models of Urban Shopping Behaviour: 4 Incorporating Independent Variables into the NBD and Dirichlet Models", Environment and Planning A 17, pp.319-331.

Yamaguchi, K. (1992) Event History Analysis. (Sage Publications, Newbury Park).

Appendix One

Fortran Program to Produce Simulated Data

```
C     Program to Produce Simulated Data Set
C
C     Written by Steven Reader and Pasquale A. Pellegrini
C
C
      PROGRAM RANFRI
      REAL A,X,Y,Q(1000),R(1000),
     $     U(528),V(528),DIS(528),CENA(528),B(460,470),
     $     CDIS(528),CENB(528),XA(5000),YA(5000)
      INTEGER IX,IY,ISEED,NR,NP,ID(528),IK(528),IK2(528)
     $       ,ID2(528),P(460,470)
      EXTERNAL RNSET,RNUN,RNGET,SSCAL,SADD
C
C
      OPEN(9,FILE='indata',STATUS='OLD')
C
C     Production of 470x460 grid to duplicate collected data
C
      X=460.5
      Y=470.5
C
C
      IONE=1
      NR=1000
      M=0
      NP=528
      A=0.5
      IX=INT(X)
      IY=INT(Y)
C
      DO 75 NA=1,NP
      CENA(NA)=1.0
      CENB(NA)=1.0
 75   CONTINUE
C
C
C
```

```
      DO 5 KY=1,IY
      DO 5 KX=1,IX
      READ(9,*) P(KX,KY)
      IF(P(KX,KY).EQ.2)THEN
      M=M+1
      XA(M)=KX
      YA(M)=KY
      ELSE
      CONTINUE
      ENDIF
 5    CONTINUE
C
C
C     Program Compares Selected Points to GIS produced study area and
C     determines if it is included or not
C
C     Compares all inter-event and boundary point Euclidean distances
C     and classifies censored and uncensored durations
C
C
      DO 85 I=1,90
C
      KA=0
      KZ=0
      N=0
C
      CALL RNSET(0)
      CALL RNUN(NR,R)
      CALL SSCAL(NR,X-A,R,1)
      CALL SADD(NR,A,R,1)
      CALL RNGET(ISEED)
      CALL RNSET(ISEED)
      CALL RNUN(NR,Q)
      CALL SSCAL(NR,Y-A,Q,1)
      CALL SADD(NR,A,Q,1)
C
      DO 10 J=1,NR
         IR=NINT(R(J))
         IQ=NINT(Q(J))
C
            IF(P(IR,IQ).NE.0.0)THEN
            N=N+1
            U(N)=IR
            V(N)=IQ
```

```
            ELSE
            CONTINUE
            ENDIF
C
C
C
            IF(N.EQ.NP)THEN
            GOTO 11
            ELSE
            CONTINUE
            ENDIF
C
C
C
 10   CONTINUE
      IF(N.LT.528)THEN
      WRITE(4,*) N
      ELSE
      CONTINUE
      ENDIF
C
C
C
 11       DMIN=9999.0
          DO 21 K1=1,NP
          DO 20 J=1,NP
          IF(J.NE.K1)THEN
          D=SQRT((ABS(U(J)-U(K1)))**2 + (ABS(V(J)-V(K1)))**2)
                IF(D.EQ.0.0)THEN
                D=1.0
                ELSE
                CONTINUE
                ENDIF
                IF(D.LT.DMIN)THEN
                DMIN=D
                DIS(K1)=DMIN
                ID(K1)=J
                ID2(K1)=J
                ELSE
                CONTINUE
                ENDIF
           ELSE
           CONTINUE
           ENDIF
```

```
 20   CONTINUE
        DMIN=9999.0
 21   CONTINUE
C
C
C
      DMIN2=9999.0
      DO 25 J=1,NP
      CDIS(J)=0.0
      CENB(J)=1.0
      DO 24 K=1,M
      D=SQRT((ABS(U(J)-XA(K)))**2 + (ABS(V(J)-YA(K)))**2)
          IF(D.EQ.0.0)THEN
          D=1.0
          ELSE
          CONTINUE
          ENDIF
          IF(D.LT.DIS(J))THEN
          ID(J)=999
          CENB(J)=0.0
C
C
C
            IF(D.LT.DMIN2)THEN
            DMIN2=D
            CDIS(J)=DMIN2
            ELSE
            CONTINUE
            ENDIF
C
C
C
          ELSE
          CONTINUE
          ENDIF
 24   CONTINUE
      DMIN2=9999.0
 25   CONTINUE
C
C     Sub-program to remove recipricol nearest neighbors
C
C
      DO 15 K=1,NP
       IK2(K)=0
```

```
      IK(K)=0
       DO 13 L=1,NP
       IF(ID2(K).EQ.L.AND.ID2(L).EQ.K.AND.L.GT.K) THEN
       IK2(K)=1
       ELSE
       CONTINUE
       ENDIF
13     CONTINUE
      IF(IK2(K).NE.1)THEN
      KZ=KZ+1
      WRITE(11,999) DIS(K),CENA(K)
      ELSE
      CONTINUE
      ENDIF
       IF(CDIS(K).EQ.0.0)THEN
          DO 12 L1=1,NP
          IF(ID(K).EQ.L1.AND.ID(L1).EQ.K.AND.L1.GT.K)THEN
          IK(K)=1
          ELSE
          CONTINUE
          ENDIF
12          CONTINUE
       IF(IK(K).NE.1)THEN
       KA=KA+1
       WRITE(8,999) DIS(K),CENA(K)
       ELSE
       CONTINUE
       ENDIF
       ELSE
       KA=KA+1
       WRITE(8,999) CDIS(K),CENB(K)
       ENDIF
15   CONTINUE
     WRITE(12,998) KZ,KA
998   FORMAT(1X,2(I3,2X))
85    CONTINUE
999   FORMAT(1X,2(F8.2,2X))
     STOP
     END
```

Appendix Two

LIMDEP Procedure to Analyze Simulated Data

```
proc
sample;n1-n2 $
surv;lhs=dur,cen;list;res=inth;fill $
sample;1-numexit $
write;dur,inth $
calc;n1=n1+x3{k1} $
calc;k1=k1+1 $
calc;n2=n2+x3{k1} $
endproc
```