Department of Systems and Information Engineering

School of Engineering and Applied Science
University of Virginia
151 Engineer's Way
P.O. Box 400747
Charlottesville, Va. 22904-4747

Final Report: Predictive Models for Law Enforcement

NIJ Grant 98-IJ-CX-K010

Donald E. Brown
Professor and Chair
Department of Systems and Information Engineering

FINAL REPORT Closeout/Archive
Approved By:
Date: 10/30/02

Introduction

Predictive Methods for Law Enforcement

Overview

Law enforcement agencies have increasingly acquired database management systems (DMBS) and geographic information systems (GIS) to support their law enforcement efforts. These agencies use such systems to monitor current crime activity and develop collaborative strategies with the local communities for combating crime. However, in general these strategies tend to be reactive rather than proactive. A more proactive approach requires early warning of trouble with sufficient lead-time to formulate a plan. Early warning, in turn, necessitates the development of predictive models in space and time that can inform law enforcement of pending "hot spots" and areas with declining crime activity.

The focus of the proposed research was on the prediction of crime events. Prediction of this sort is now feasible because of modern data collection and analysis systems. Records management systems implemented in DBMS and GIS exist in many jurisdictions and can provide the basis for more formal analysis of local crime events. The formal analysis that we developed consists of mathematical models that describe the functional relationships between demographic, economic, social, victim, and spatial variables and numerous measures of criminal activity.

The results presented in this report show impressive effectiveness in predicting crime. Despite coarse feature variables, the method outperformed standard density estimation techniques for identifying regions of increased crime activity. We believe this the approach developed through this grant provides promise for both more accurate prediction of criminal events as well as testing theories regarding factors that contribute to rising crime rates.

Objectives

The goal of this research was to develop predictive models that would enable law enforcement agencies and their communities to proactively address criminal activity. We had the following specific research objectives.

1. Devise and implement predictive models that specifically address the needs of law enforcement.
2. Investigate and implement methods that can identify the most useful features for criminal incident prediction.
3. Empirically evaluate the effectiveness of the prediction models and feature selection techniques using data from Richmond and/or Charlottesville-Albemarle.
4. Disseminate the most promising of the models for use by law enforcement agencies.

Contents

This report provides results from our research organized into three papers. The first of these provides the theoretical foundation for our new approach to crime event prediction. This approach is built out of results in space-time point processes. We review fundamentals from this area and then give the theoretical details of our approach.

The second paper shows applications of our approach to a crime prediction problem in Richmond, Virginia. Specifically, we examined breaking and entering events and used data from one week to predict both the next week and the next two weeks. We compared our predictions to those provided by several density estimation approaches (e.g., kernel estimates) and found that we significantly outperformed these estimates. This suggests that the use of feature data can improve the predications of criminal events. We believe that even better (more accurate, higher resolution) feature data will further improve performance.

The third paper and final paper provides an extension of the model to handle temporal features. The paper discusses the problem of measuring similarity between temporal features and shows our approach for handling this problem. We then show how temporal features can be used with point process model to provide for both space-time attributes. We tested this approach using data from Richmond, Virginia and found that in some cases the temporal features improved performance, but this was not always the case. Clearly to use these features effectively, we need a filter that identifies incidents that have low variance in certain temporal features. These incidents then become the best candidates to use temporal features in prediction.

The last paper also provides information on our implementation of the prediction methodology. In to order to complete this research we had to build a more flexible interface into the algorithms underlying the methodology. The papers shows some of the results from this work since they could serve as the foundation for implementing the approach in crime analysis software.

Theoretical Foundations for a New Point Process Transition Density Model for Space-Time Event Prediction

Hua Liu
Lucent Technologies, Inc.
One Main Street
Cambridge, MA 02142
hualiu@lucent.com

Donald E. Brown
Department of Systems Engineering
University of Virginia
Charlottesville, VA 22903
brown@virginia.edu

Abstract: A new point process transition density model is proposed based on the theory of point patterns for predicting the likelihood of occurrence of spatial-temporal random events. The model provides a framework for discovering and incorporating event initiation preferences in terms of clusters of feature values. Components of the proposed model are specified taking into account additional behavioral assumptions such as the "journey to event" and "lingering period to resume act." Various feature selection techniques are presented in conjunction with the proposed model. Extending knowledge discovery into feature space allows for extrapolation beyond spatial or temporal continuity and is proved to be a major advantage of our model over traditional approaches. We examine the proposed model in the context of predicting criminal events in space and time.

Key Words: Space-Time Marked Shock Processes, Forecasting, Criminal Event Prediction, Probability Density Estimation

1. Introduction

Consider the following scenario from the domain of law enforcement: Within a monitoring region marked by jurisdictional boundaries, a crime analyst is interested in mapping out the areas that are more likely to be struck by a certain type of crime within a given time range. Data available to the analyst are the past crime incidents of the same type, times of occurrence, locations of occurrence, and characteristics (or features) of the crime scene. The problem facing the crime analyst is how to extrapolate these data into the likelihoods of future incidents occurring at specified locations in space and time. Ideally the analyst wants an image map showing the intensities of future crime activities at each location within the their jurisdictional boundaries. This solution and its display

are obviously useful in crime prevention since they would allow the police to allocate the police resources to the areas of higher risk.

This problem is not confined to law enforcement. For example, in military actions, one may want to predict the future location of an enemy target (e.g., a tank) moving over terrain based on its past locations (observed over predefined sampling intervals) and terrain features. In an urban development, developers are interested in predicting consumer behavior toward a new shopping mall using data from past behavior toward existing malls. They would also use data regarding surrounding neighborhoods and the physical infrastructure in the area (e.g., major highways, schools, and bridges).

The common characteristic in these problems is prediction based on spatio-temporal event data. A number of researchers have investigated forecasting over space and time. A significant advance in this area was the development of space-time autoregressive moving average (STARMA) models [10]. These models offer a way of generalizing the ARMA (autoregressive moving average) models in time series analysis to combined spatial-temporal domains. They are characterized by linear dependence lagged in both space and time. Several authors [1], [3], [26], [32] discussed issues of stationarity and invertibility arising from model parameter estimation based on the assumption that spatial dependence is instantaneous. The STARMA model was further generalized by Pfeifer and Deutsch [33] to include temporal differences (STARIMA models) and by Stoffer [41], [42] to include a nonstationary mean function of independent variables (STARMAX models).

While the STARMA models may effectively and simultaneously capture the continuity in space and time, they fail to take into account event-related feature

information which may very well reveal and represent the underlying pattern of event occurrences. This is especially important when event initiation is marked by human intelligence. We argue that for "intelligent" human initiated events, their future locations are correlated to a larger extent with site selection preferences (as represented by event-related features) than with spatial proximity. By extending analysis into feature space, we are able to identify highly likely future event locations that are not necessarily in the vicinity of past event locations. This is the very aspect where the STARMA models fail.

The central theme of this paper is to develop a new space-time prediction model that incorporates all three kinds of event data (i.e., times, locations, and features). In particular, the available data are viewed as a realization of a marked space-time shock point process, and the space-time prediction problem is formulated as the estimation of the transition density of the stochastic process (Section 2). A model of the transition density is constructed and the underlying technical assumptions are justified by the behavioral theories of events' initiation (Section 3). Next, we discuss the criteria for selecting key features (Section 4) and the procedures for estimating individual components of the proposed model (Section 5). Finally, we summarize the advantages of our model and point out some future research directions (Section 6). Throughout this paper, we use criminal event prediction as a motivating example.

2. Problem Statement

The space-time prediction problem we described in last section can be stated as follows: Having observed a series of events of the same type (e.g., incidents of a type of crime) in a monitoring region, namely, the locations and times of the events, and the values of an array of features that are known or believed to be relevant to the occurrence

of the events, we would like to predict the likelihood that another event occurs at certain location within the region and within a certain time range. Mathematically, we consider the locations ($\mathbf{s}_i$) and times (t_i) of the events, $(\mathbf{s}_1,t_1)$, $(\mathbf{s}_2,t_2)$, ..., $(\mathbf{s}_n,t_n)$, $t_0 = 0 < t_1 < t_2 < ... < t_n$, and their corresponding features (or marks), $\mathbf{x}_{\mathbf{s}_1,t_1}$, $\mathbf{x}_{\mathbf{s}_2,t_2}$, ..., $\mathbf{x}_{\mathbf{s}_n,t_n}$, as **a realization** of a *marked space-time shock point process* of the form

$$\left\{\mathbf{x}_{\mathbf{s},t} \in \chi : \mathbf{s} \in D, t \in T\right\} \tag{2.1}$$

where t, $\mathbf{s}$, and $\mathbf{x}_{\mathbf{s},t}$ are all random (bold indicates vectors). The location of an event is confined within a *study region* or *geographic space* $D \subset \Re^2$ and is designated by a pair of coordinates, say longitude and latitude, i.e., $\mathbf{s} = (s_1, s_2)$. $T \subset \Re^+$ is the collection of the times when the events could occur, and is termed the *study horizon*. $\chi \subset \Re^p$ is the collection of the possible values of the p-dimensional feature vectors (i.e., the marks), and is termed *feature space*. Let $F^{(p)} = \{f_1, f_2, ..., f_p\}$ where f_l, $l = 1,2,...,p$, is the lth feature or the lth dimension of the feature space. Then each $\mathbf{x}_{\mathbf{s}_i,t_i}$ is an instantiation of these p features. We abbreviate $\mathbf{x}_{\mathbf{s}_i,t_i}$ as $\mathbf{x}_i$ from now on. But the reader should bear in mind that $\mathbf{x}_i$'s are feature observations of different events and taken together these events comprise one realization of the point process. Let $T_n = \{t_1, t_2, ..., t_n\}$, $D_n = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ and $\chi_n = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{s}_i = (s_{i1}, s_{i2})$ and $\mathbf{x}_i = [x_{i1} \quad ... \quad x_{ip}]'$. Given that we observed T_n, D_n, and χ_n up to instant t_n, we are interested in estimating, for $\mathbf{s}_{n+1} \in D$ and $t_{n+1} > t_n$, the transition density

$$\psi_n(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n, \chi_n) = \lim_{\nu(d\mathbf{s}_{n+1}).dt_{n+1} \to 0} \frac{\Pr\{N(d\mathbf{s}_{n+1}, dt_{n+1}) = 1 | D_n, T_n, \chi_n\}}{\nu(d\mathbf{s}_{n+1}) dt_{n+1}} \tag{2.2}$$

where $\mathbf{s}_{n+1}$ and t_{n+1} are the realizations of the location and the time of the next event, respectively, $\nu(d\mathbf{s}_{n+1})$ is the Lebesgue measure of $d\mathbf{s}_{n+1}$ and $N(d\mathbf{s}_{n+1}, dt_{n+1})$ counts the incidents that happen within $d\mathbf{s}_{n+1}$ and dt_{n+1}.

By (2.2), the transition density is formally defined as the probability that a single event occurs within a specified infinitesimal region (e.g., $d\mathbf{s}_{n+1}$) and within a specified infinitesimal time interval (e.g., dt_{n+1}). "Single" or uniquely identifiable events are ensured in theory if we postulate a *simple* point process. In practice, however, one should pay attention to what constitutes single events. The notion has no bearing on either event scale or event duration. For example, the Oklahoma City Bombing involving massive explosions and multiple casualties and bombing of an abortion clinic with a single explosion and no serious injuries are both considered single bombing incidents since they can both be uniquely identified by the unique location and time of occurrence.

Prior to model development, we require additional description of the set of "features". First, we divide the set of features into the set of (*inherently*) *temporal features* and that of all others. By "(inherently) temporal features", we mean features that "label" time intervals so that categorization of time instants can be obtained. Some examples are "seasons of the year", "holiday / non-holiday", "segments of a day (e.g., morning / afternoon / night)". A temporal feature partitions the time axis $\Re^+$ into consecutive time intervals, and the time instants in a single interval are all identified with the same temporal category. The purpose of segmenting the time axis with a temporal feature is to provide us with suitable and meaningful time intervals in which we may postulate stationary models for the temporal aspect of the process. Depending on which temporal category t_{n+1} belongs to, we may only use (local) data in the same category in

these models. Modeling temporal heterogeneity (i.e., heterogeneity in the series $t_1, t_2, ..., t_n$) is not the focus of this work. However, theoretically we can incorporate temporal heterogeneity by using ARIMA-like models for temporal transition estimation. This will add a common factor for every location in the study region at any given time with the transition density model we are proposing (discussed further in the next section). Since the synthesized effect of different temporal categories on event occurrence is contained in the complete series $t_1, t_2, ..., t_n$, we may exclude all temporal features from the feature set $F^{(p)} = \{f_1, f_2, ..., f_p\}$. Additionally we assume temporal features are independent of geographic locations. Formally, we make the following assumption:

Assumption 2.1: $F^{(p)} = \{f_1, f_2, ..., f_p\}$ is the set of features that depend on (but not necessarily only on) geographic locations; i.e., the feature space $\chi \subset \Re^p$ does not contain temporal features.

Secondly, although many features (e.g., proximity to major highways) can be considered static within the study horizon if we do not consider randomness involved in taking the measurements, we nevertheless model the feature vector in its entirety as random to take into account features of probabilistic nature (e.g., occupancy of the victimized household, race of the offender). For static features, we regard them as random variables taking on certain values with probability one. Static features are usually directly derived from relations with static geographic surroundings or establishments.

3. Model Development

We describe a model that captures the mechanism governing event occurrences over the study horizon and the study region. Model development consists of a two-step decomposition of the transition density $\psi_n(\mathbf{s}_{n+1}, t_{n+1}|D_n, T_n, \chi_n)$. In this Section, we combine both intuitive and formal descriptions of our model.

The first step of the decomposition is to separate the spatial and temporal transitions. We postulate that the occurrences of events over time and space are separable in the sense that

$$\psi_n(\mathbf{s}_{n+1}, t_{n+1}|D_n, T_n, \chi_n) = \psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1}) \cdot \psi_n^{(2)}(t_{n+1}|T_n) \qquad (3.1)$$

where $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1})$ will be called *spatial transition density* and $\psi_n^{(2)}(t_{n+1}|T_n)$ *temporal transition density*. Equation (3.1) would be a standard Bayesian decomposition if the second term on the right-hand side was $\psi_n^{(2)}(t_{n+1}|D_n, \chi_n, T_n)$. D_n and χ_n were left out under two assumptions: Assumption 2.1 specified in last section and Assumption 3.1 as follows.

Assumption 3.1: Temporal evolution (transition) of the point process (2.1) does not depend on spatial (locational) evolution (transition).

In other words, we assume that spatial dependence arises from the integration of causal factors over time, but not vice versa. In the crime analysis scenario, for example, we do not regard the past crime intensity at a site as a direct factor to influence how soon criminals are going to strike again. However, this past behavior does tell us about the preferences of site selectors and we directly model these preferences in the second step of the decomposition below.

We now proceed with this second step of the decomposition: modeling the spatial transition density $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1})$. To develop this model, we first introduce some behavioral theory that accounts for the intelligent site selection by event initiators, and then give the relationship between features of geographic locations and site selection behavior. Intuitively speaking, our modeling philosophy is to use past site selection behavior to inform where events are likely to occur again.

For human-initiated events over a geographic region, one primary behavioral assumption is that *event initiators (e.g., offenders in crime scenario) choose the site of an event based upon a set of preferences over the values of the attributes (features) at alternative sites*. This is well documented for the crime analysis scenario as it appears frequently in criminology literature [8], [30], [31], [35], [38]. Preferences are measured in feature space (χ), and a *set of preferences (pertaining to a group of event initiators)* is defined when a subset of features (corresponding to the set of spatial attributes **actually** considered by the group of event initiators) and a partial ordering of available values for these features are specified. For a specific group of event initiators, if we knew their set of preferences (i.e., the subset of features and the partial order for each feature), we would examine all locations in geographic space for their feature values and score them accordingly. However, without this knowledge of site selection preferences, we must "discover" it from the data, or specifically, from the point pattern in feature space. We make two assumptions here: (1) *If multiple groups of event initiators are present, we assume that they make site selection decisions based on common set of features*. This assumption is inevitable if we want to deal with multiple groups simultaneously. (2) *The set $F^{(p)}$ of features that we choose initially coincides with that of the event initiators (the*

"true" feature set). By making this assumption, we postpone part of the knowledge discovery task (feature selection) until the next section. To establish the relationships between site selection preferences and the point pattern in feature space, we essentially rely on this "stationarity" assumption: *Preferences remain stable (stationary in a probabilistic sense) over the study region and study horizon for each group of event initiators.* Given the data of repeated site selection decisions by a group, the set of preferences of this specific group must manifest itself as a small-variation distribution of values in feature space. This small-variation distribution can be described as a *clique* in point process theory (or less formally as a *cluster*). If multiple groups with distinct preferences are present over the study region, we expect to see a clustering (point) pattern with multiple cliques in feature space (See Figure 3.1).

The second behavioral assumption for intelligent site selection is concerned with the spatial interaction or dependence between selected sites over the study region. Given that two geographic locations have the same set of feature values, it is often reasonable to postulate that *event initiators are in favor of the geographically closer location to start the next event.* For example, the "journey to crime" theory in criminology states that the distance to the place of the crime is important [2], [5], [9], [22] and many types of crimes have their own defined "radii" [36], [37]. In view of this assumption, a model of spatial interaction should give decreasing weight to past events with increasing distance to the location of interest. Another behavioral assumption that may hold true for certain scenarios (e.g., serial crimes of certain type) is that *event initiators tend not to wait long before they act again.* A model incorporating this assumption should weigh the impacts

of past events on future events according to their "ages". The more recently an event occurred, the higher weight it gets.

Figure 3.1 illustrates event occurrences in different spaces. Although the distribution of events on time axis as well as that in geographic space could very much

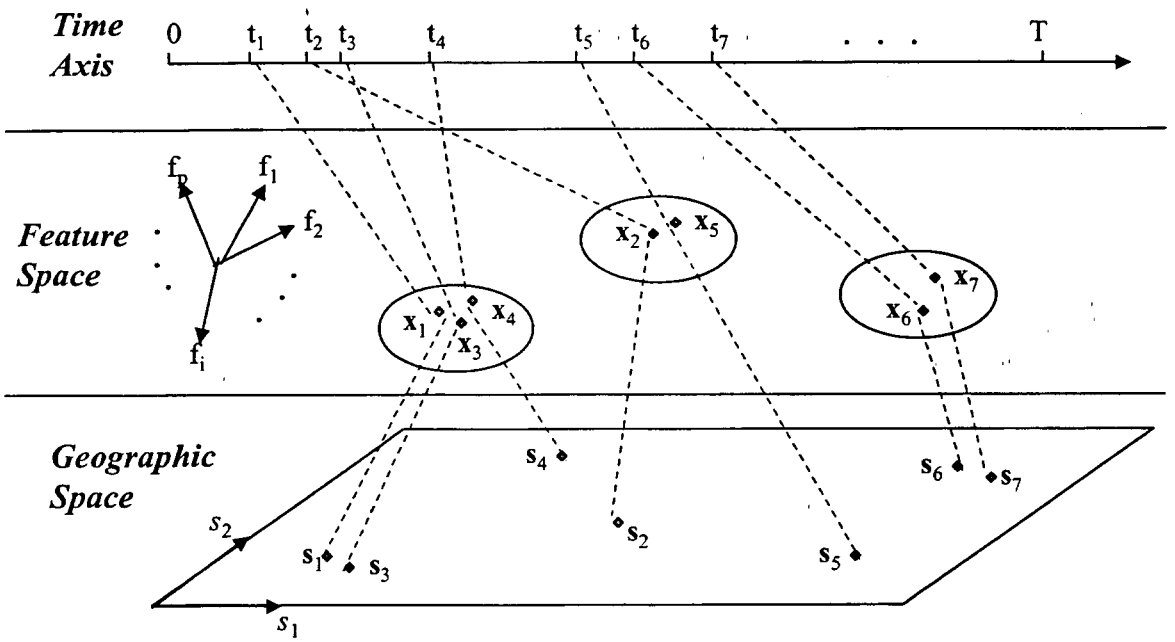

Figure 3.1. Three views of event occurrences.

lack any systematic pattern, stable and distinct clustering patterns should be observed in feature space. Each clique in feature space corresponds to a set of preferences. It is often the case that locations in close geographic proximit have similar feature values. Then neighbors in geographic space are neighbors in feature space (e.g., $\mathbf{s}_6$ and $\mathbf{s}_7$). However, proximity in feature space does not necessarily translate into proximity in the geographic space (e.g., $\mathbf{s}_2$ and $\mathbf{s}_5$). It is quite possible that two locations that are far apart have the same feature values and thus it is only reasonable to assign an equal score to both locations if we extrapolate event occurrence based solely upon site selection preferences.

The merit of integrating feature space information into space-time event prediction is that **potential** event areas (e.g., areas not previously struck as frequently by crimes but at high risk nevertheless) can be picked out.

We are ready to formally describe our model for spatial transition density. Suppose that the set χ_n of feature vectors is partitioned into C disjoint subsets $\{\chi_n^{(j)} : j = 1,2,...,C\}$ corresponding to the cliques in feature space (i.e., sets of preferences). Correspondingly, the set D_n (T_n) of locations (times) of past events is also partitioned into C disjoint subsets $\{D_n^{(j)} : j = 1,2,...,C\}$ ($\{T_n^{(j)} : j = 1,2,...,C\}$). Let $\mathbf{x}_{n+1}$ be the estimated feature values at location $\mathbf{s}_{n+1}$ and instant t_{n+1}. Conditional on $\mathbf{x}_{n+1}$, the transition density $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1})$ in (3.1) is assumed to take the form

$$\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1}) = \alpha \cdot \psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n) \cdot \sum_{j=1}^{C} \psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1}) \Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\} \tag{3.2}$$

where $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$ is termed the *first order spatial transition density** and reflects event intensity (i.e., first order effects) at $\mathbf{x}_{n+1}$ in feature space. $\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1})$, $j = 1,2,...,C$, are termed *second order spatial transition densities*, which reflect interaction (i.e., second order effects) of new event location $\mathbf{s}_{n+1}$ with past event locations in each $D_n^{(j)}$, respectively. $\Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\}$, $j = 1,2,...,C$, are *spatial interaction probabilities* or the probabilities that $\mathbf{x}_{n+1}$ and each $\chi_n^{(j)}$ form a clique in the feature space. α is a normalizing constant.

* This is probability mass function in the case of discrete feature space. We shall use the term "density" in both continuous and discrete cases.

Model (3.2) incorporates all elements of site selection behavior and puts them into a formal framework — spatial point process theory. A spatial point pattern can be regarded as the result of first order effects coupled with second order effects. We model first order effects as the event initiators' site selection preferences or alternative sites' potential to attract future events (**feature space** analysis) rather than the average number of events already accumulated at alternative sites (geographic space analysis). This notion of site selection preferences is more fitting for prediction given that the same sets of preferences will carry on to t_{n+1} over the study region ("stationarity" assumption). Technically, the assumptions concerning site selection preferences can be considered equivalent to the following assumption:

Assumption 3.2: The spatial point process in (true) feature space is Markovian over a small range.

Roughly speaking, this assumption ensures that in feature space, there are no second order effects (i.e., dependence or interaction) between cliques, and since range (or clique radius) is small, only first order effects are important within each clique. In correspondence with the behavioral assumptions concerning spatial dependence, the second order effects are modeled in **geographic space**. Notice that it is only appropriate to examine spatial dependence for events in the same feature-space clique (i.e., events initiated by the same group of people). However, due to the uncertainty associated with assigning a new event to a specific clique (or claiming that a specific group is responsible for a new event), we weigh second order effects pertaining to individual cliques by the probabilities that they quantify the uncertainty (i.e., spatial interaction probabilities). Technically, we estimate the weighted average of the second order effects of C thinned

point processes in geographic space, aiming to maintain continuity in parallel with the ordering of inter-event geographic distances and/or that of inter-event temporal distances. A realization of each thinned point process is the set $D_n^{(j)}$ of events corresponding to those that form the clique $\chi_n^{(j)}$ in feature space.

Finally, we need to point out that the spatial transition density model (3.2) needs "prior" adjustment when the predicted feature values ($\mathbf{x}_{n+1}$'s) for all locations within the study region (D) do not form a uniform distribution. Let $\kappa_n(\mathbf{x}_{n+1})$ denote the probability density function of $\mathbf{x}_{n+1}$ over all predicted feature values for locations $\mathbf{s}_{n+1} \in D$. Non-uniformity of $\kappa_n(\mathbf{x}_{n+1})$ indicates certain feature values are more typical than others in the study region. Individual locations with typical feature values, if preferred by event initiators, should be at lower risk compared with those with rare feature values simply because event initiators have more choices over the region but they may engage themselves at only one location at any instant*. To put all locations on an equal footing, we adjust (3.2) as follows.

$$\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n,\chi_n,T_n,t_{n+1}) = \beta\cdot(1/\kappa_n(\mathbf{x}_{n+1}))\cdot\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$$
$$\cdot\sum_{j=1}^{C}\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)},T_n^{(j)},t_{n+1})\Pr\{\mathbf{x}_{n+1}\in\chi_n^{(j)}\} \qquad (3.3)$$

where β is a normalizing constant. When $\kappa_n(\mathbf{x}_{n+1})$ is uniform, (3.3) reduces to (3.2). $\kappa_n(\mathbf{x}_{n+1})$ can be easily estimated in the case that all features are static over the study horizon. We use (3.2) when we do not have knowledge of $\kappa_n(\mathbf{x}_{n+1})$. We term $\kappa_n(\mathbf{x}_{n+1})$ the *geographic-space feature density*.

---

* Technically, we have assumed that no two events happen at the same location or at the same time.

In order to implement the spatial-temporal transition density model given by (3.1), (3.2) and (3.3), two areas of work need to be done. They are selection of the *key features* to be used in the model and estimation of individual model components. We address these two areas in the next two sections, respectively.

4. Feature Selection

For real applications, we frequently come across a fairly large initial feature set $F^{(p)}$. Large amounts of data are good for us in the sense that we have a better chance of covering the set of spatial features that actually prompt the selection of past event locations, or the true feature set. However, it is also natural to conclude that not all features in the initial set carry equal weights towards event initiation. In fact, we want to *find the smallest feature subset (of the initial feature set $F^{(p)}$) that is necessary and sufficient to account for the underlying spatial pattern of event occurrences.* A small or parsimonious feature subset is important for building an empirical model. It has been long understood that an empirical model constructed with a larger number of features may fit the training data set quite well but it seldom generalizes nearly as well on new data sets. We term the selected feature subset the *key feature set* and denote it as $F^{(q)}$, where q is number of *key features* contained in $F^{(q)}$ and $1 \le q \le p$. The feature subspace defined by $F^{(q)}$ is termed the *key feature space.*

A feature selection problem can generally be specified by a triplet (F, c, s), where F is the *initial feature set*, c a *criterion function* defined for subsets of F, and s is a *subset search or selection procedure.* Our emphasis in this paper is on feature selection criteria. In particular, we discuss briefly the unique characteristic of the feature selection problem

in the intelligent event initiation scenario and then summarize two categories of feature selection criteria applicable to the problem. In theory, a number of exact or inexact feature selection procedures may be used with these criteria to identify the key feature set.

Feature selection problems have been discussed mainly within the contexts of two research areas: pattern recognition and regression model building. In either area, there is a target concept (as defined by a class variable in pattern recognition and a response variable in regression) "outside" of the features. A natural "external" feature selection criterion would be how well the outside target is described by the chosen feature subset. Unlike these areas, there is not an "outside" target concept in the intelligent event initiation scenario. All information regarding event initiation preferences is contained in the feature data. In other words, the feature data are "unsupervised". In this case, we need to resort to "internal" criteria, which describe the desired (or intrinsic) data structure (or point pattern) in the lower-dimensional feature space believed to be the true feature space. The intrinsic data structure is usually problem-specific and only partially observable in the initial feature space.

Based upon our analysis in the last section, a distinct clustering pattern consisting of small and well-separated cliques should be observed in the true feature space. Consequently, it is required that, as the first and primary criterion, the events form a distinct clustering pattern or a *cohesive* point pattern in the key feature space. All the initial features were chosen with the belief that they are all somewhat relevant to event occurrence. Therefore, as a second criterion, if any systematic pattern is manifest in the initial feature space, it should be roughly preserved in the key feature space. We say

"roughly" due to two facts. First, the initial pattern or data structure cannot be completely preserved in any feature subspace for real problems. Second, we may in fact only wish to roughly preserve the initial pattern since some initial features are unimportant to the event initiators. We give the technical details in the subsections below.

4.1. Measures of Cohesiveness of a Point Pattern

A point pattern or data (distributional) structure is intuitively said to exhibit *cohesiveness* if it consists of distinct clusters (or cliques), i.e., clusters that are tight individually and well separated between one another. Cohesiveness can be gauged in many ways. One class of measures assumes that the data are already "optimally" partitioned into the "best number" of clusters using some clustering algorithm. The measures in this class are essentially functions of the cluster means and the cluster covariances. We may construct a measure of this class based on the Mahalanobis D^2. Let C be the number of clusters in the partition. The Mahalanobis D^2 for a pair of clusters is the squared distance between the means of the two clusters normalized by the pooled covariance matrices.

$$D_{ij}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)'(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{4.1}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean vector and covariance matrix of the events in an individual cluster. The one-dimensional pairwise Mahalanobis D^2 is known to statisticians as the Fisher ratio. The averaged Mahalanobis D^2 score is given by

$$mad2 = \frac{2\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\pi_i\pi_j D_{ij}^2}{C(C-1)\sum_{i=1}^{C-1}\sum_{j=i+1}^{C}\pi_i\pi_j} \tag{4.2}$$

where π_i and π_j are *a priori* probabilities of clusters i and j. We may normally let $\pi_i = n_i/n$, where n_i is the number of events in cluster i, when no further *a priori*

information is available. Apparently, we can expect large inter-cluster separation and small intra-cluster spread when *mad2* is maximal. We may transform *mad2* into a measure ranging between 0 and 1 as follows.

$$I_m = \frac{1}{1 + mad2} \tag{4.3}$$

The transformed measure I_m is to be minimized. Friedman and Rubin [17] constructed several simpler criteria in the same class. Instead of looking at the clusters in pairs, they used a "pooled within-groups scatter matrix" and a "between-groups scatter matrix" constructed from all clusters. One practical problem for this class of the measures is that we have need the "optimal" partition with the "best number" of clusters to actually evaluate these measures. What is meant by "optimal" is obviously dependent upon the clustering algorithm used. Besides, it is no trivial problem just to determine the "best number" of clusters, or equivalently whether an event is an "outlier", unless this can be specified by *a priori* knowledge.

Another class of measures of cohesiveness does not require any partitioning in advance. These measures are functions of inter-event distances (or similarities). No matter what functional forms are adopted, these measures always account for a basic characteristic of cohesive structures: It is nearly always the case that the distance between a pair of events is either very small or very large (i.e., the two events are either very similar or very dissimilar); rarely are two events separated by average inter-event distance. Dash et al. [11] used an entropy-based measure. For two events i and j, and similarity s_{ij}, the entropy measure is defined as

$$e_{ij} = -s_{ij} \log_2 s_{ij} - (1 - s_{ij}) \log_2 (1 - s_{ij}) . \tag{4.4}$$

This is a bell-shaped curve that assumes its maximum value of 1.0 for $s_{ij} = 0.5$, and the minimum value of 0.0 for $s_{ij} = 0.0$ and $s_{ij} = 1.0$. The entropy-based measure of cohesiveness for a data set of n events was given as

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij}, \tag{4.5}$$

which attains its minimum value when the data structure (or point pattern) is most cohesive. We prefer the normalized form

$$I_e = \frac{2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e_{ij}}{n(n-1)}. \tag{4.6}$$

The rationale of the entropy-based measure is as follows. For a pair of events, consider the binary random variable that has two opposite outcomes: The two events belong to the same cluster or they do not. When two events are more similar, they are more likely to be in the same cluster. Informally, let us assume that the similarity of the two events is equal to the probability that the two events are in one cluster. From information theory, when the outcome of the binary variable is most uncertain (i.e., both outcomes have equal probabilities of occurrence), the entropy (as defined by (4.4)) is the largest. If for most pairs of events, we cannot conclude with much certainty whether they belong to the same cluster or not, (4.6) is maximized. Therefore, a minimal value of the entropy-based measure corresponds to a point pattern that exhibits cohesiveness.

We submit that any other bell-shaped functions may be used in place of (4.4) to effectively account for the aforementioned characteristic of cohesive patterns. Renyi's entropies [6], [34] and the Gini index of diversity are among a number of useful choices. We used a measure based on the Gini index in the following due to its simple form. For a pair of events i and j, the Gini index is defined as

$$g_{ij} = 4s_{ij}(1-s_{ij}). \tag{4.7}$$

Like e_{ij}, g_{ij} attains its maximum of 1.0 at $s_{ij} = 0.5$ and its minimum of 0.0 at $s_{ij} = 0.0$ and $s_{ij} = 1.0$. For a data set of n events, the averaged Gini index (4.8) is a suitable measure of cohesiveness. Again, smaller I_g corresponds to higher level of cohesiveness.

$$I_g = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} g_{ij}}{n(n-1)}. \tag{4.8}$$

To illustrate these ideas, we consider a simple example. Figure 4.1 shows two of a series of 1-dimensional 4-event point patterns within a fixed range, where b is the distance between the two events at either end and c is the distance between the two events in the middle. For 4.2(a), we observe two clusters with b measuring within-cluster scatter and c between-cluster scatter. As the ratio b/c increases, b and c switch roles and we observe a three-cluster pattern like 4.2(b), with the two events in the middle falling into one cluster.

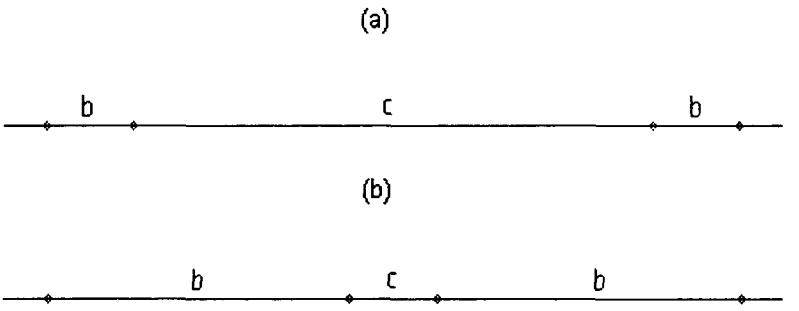

Figure 4.1. 1-D 4-event point patterns within a fixed range.

Transforming inter-event distance d_{ij} into similarity s_{ij} using

$$s_{ij} = \frac{1}{1+d_{ij}/\bar{d}}, \tag{4.9}$$

where $\bar{d}$ is the mean of all inter-event distances, we have

$$I_g = \frac{4+6z}{(5+3z)^2} + \frac{8+20z+12z^2}{(5+6z)^2} + \frac{4+14z+12z^2}{(5+9z)^2} + \frac{2+3z^2}{(1+3z)^2}, \qquad (4.10)$$

where z denotes b/c. This is plotted in Figure 4.2.

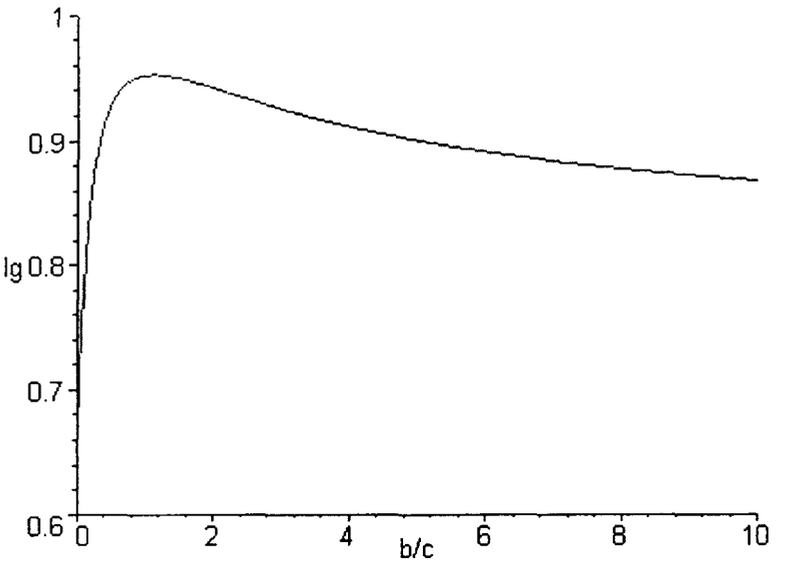

Figure 4.2. Gini index as a measure of cohesiveness.

Not surprisingly, the minimal I_g occurs when $b/c = 0$, which corresponds to the most cohesive two-cluster pattern (with each cluster having two events) within the fixed range. I_g increases dramatically when b/c increases and the four events become more evenly spaced within the range. It reaches its maximum at around 1.117 or $b \approx c$. Then it decreases gradually as the events group into three clusters. But even when the most cohesive three-cluster pattern is observed (i.e., when $b/c \to \infty$), I_g never drops lower than its value when $b/c = 0$. This appeals to our intuition: Spanning the same range, two-cluster patterns appear more cohesive than three-cluster ones.

Notice that none of the three aforementioned measures are intended for addressing a single-cluster structure. They are calculated only if there is enough variation on every dimension of the data set for feature evaluation (relative to the full range of that

dimension over the entire region of interest). Operationally, we check each dimension of the data set and exclude the features that do not exhibit enough variation. Domain knowledge needs to be exercised to determine whether these features are the most predictive ones or the most irrelevant ones to the problem at hand. Technically, zero-variation features need to be singled out anyway to avoid singularity when fitting density estimation models.

4.2. Measures of Disagreement between Point Patterns

Another criterion for dimensionality reduction of unsupervised data stems from the observation that the intrinsic data structure or point pattern should be partially observable in the initial feature space due to the fact that all initial features are believed to be somewhat predictive. The idea is illustrated in Figure 4.3. Suppose that we initially have

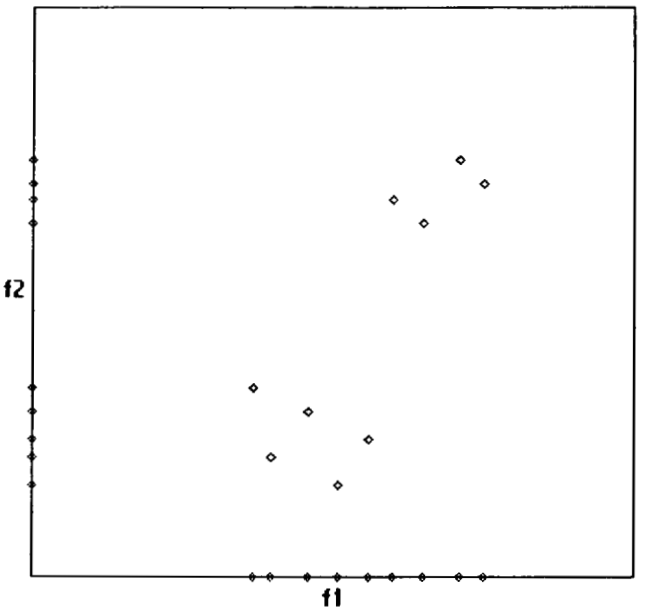

Figure 4.3. A 2-D 9-event point pattern and its 1-D projections.

a 9-event point pattern in two-dimensional space. Most people observe two clusters in the two-dimensional space. The events, when projected onto the f_1-axis, all lump together, whereas the two-cluster pattern can still be clearly identified if projected onto f_2-axis. We say that feature f_2 is more important than feature f_1 since the original pattern is preserved on the f_2-axis. Clearly, to compare different projections, we need a measure of disagreement to quantify how much difference exists between the initial point pattern and a lower-dimensional projection.

The measures of disagreement between point patterns fall under two classes. The first class, again, assumes that the data set have been partitioned in the *best* way. The task then is to quantify the disagreement between the partition obtained with all initial features included and that obtained with selected features. Birkenhead [7] gave several measures in this class. We adapt the simplest one of these for our use here. Suppose that the partition in a feature subspace consists of C clusters, $\varpi_1, \varpi_2, ..., \varpi_C$, discovered by applying some clustering algorithm. The same clustering algorithm reports C' clusters, $\Omega_1, \Omega_2, ..., \Omega_{C'}$ in the initial feature space. For a pair of instances i and j, define a score

$$\begin{aligned} b_{ij} &= 0 \qquad \text{if } i, j \in \varpi_u \text{ and } i, j \in \Omega_v \text{ for some } u \text{ and } v, \\ b_{ij} &= 0 \qquad \text{if } i, j \notin \varpi_u \text{ and } i, j \notin \Omega_v \text{ for any } u \text{ and } v, \\ b_{ij} &= 1 \qquad \text{otherwise.} \end{aligned} \tag{4.11}$$

The disagreement between the point patterns in initial feature space and the feature subspace is then set as

$$I_b = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} b_{ij}}{n(n-1)}. \tag{4.12}$$

Notice that $0 \le I_b \le 1$ with $I_b = 0$ indicates the least disagreement or identical point patterns. $I_b = 1$ when the two point patterns are $\{\{1,2,...,n\}\}$ (all instances in one cluster) and $\{\{1\},\{2\},...,\{n\}\}$ (every instance in its own cluster). It can be shown that this disagreement measure actually satisfies the properties of a metric or distance measure [7].

Besides I_b, we have found the following normalized divergence measure also works reasonably well. Define the conditional probabilities

$$\Pr(\varpi_u|\Omega_v) = \frac{|\varpi_u \cap \Omega_v|}{|\Omega_v|} \quad \text{and} \quad \Pr(\Omega_v|\varpi_u) = \frac{|\varpi_u \cap \Omega_v|}{|\varpi_u|} \tag{4.13}$$

for $u = 1,2,...,C$ and $v = 1,2,...,C'$. Let P_ϖ and P_Ω denote the partitions $\{\varpi_1, \varpi_2, ..., \varpi_C\}$ and $\{\Omega_1, \Omega_2, ..., \Omega_{C'}\}$, respectively. Define the divergence score

$$div(P_\varpi, P_\Omega) = \left(\sum_{u=1}^{C}\sum_{v=1}^{C'}\left[\Pr(\varpi_u|\Omega_v) - \Pr(\Omega_v|\varpi_u)\right]\log_2\left[\Pr(\varpi_u|\Omega_v)/\Pr(\Omega_v|\varpi_u)\right]\right)/CC' . \tag{4.14}$$

This score attains its minimum, 0, when the two partitions P_ϖ and P_Ω are identical. It is easily shown that

$$\max_{P_\varpi, P_\Omega}\{div(P_\varpi, P_\Omega)\} = div(\{1,2,...,n\},\{\{1\},\{2\},...,\{n\}\}) = \frac{n-1}{n}\log_2 n . \tag{4.15}$$

The normalized divergence measure is given as

$$I_d = \frac{div(P_\varpi, P_\Omega)}{\max_{P_\varpi, P_\Omega}\{div(P_\varpi, P_\Omega)\}} \tag{4.16}$$

Again, this class of disagreement measures depends upon the clustering algorithm used for partitioning.

The second class of measures for comparing point patterns comes from the area of multidimensional scaling (see [20], [21], [40] for a review of early work). This approach does not require that the data are already partitioned. Multidimensional scaling deals with

the following problem: For a set of observed distances (or similarities) between every pair of n events, find a representation of the data in fewer dimensions such that the inter-event distances (or similarities) as they are measured in the subspace "nearly match" the initial similarities (or distances). Multidimensional scaling is only done with *ordinal* or *nonmetric* data, i.e., the rank orders of the $n(n\text{-}1)/2$ of inter-event distances*. This is because a particular rank ordering corresponds to a set of point patterns whose geometrical difference or disagreement is quite small. For example, for a two dimensional pattern containing as few as 20 instances, a movement of only one instance on the order of 0.1% of the largest inter-instance distance in the pattern will result in a modification of the rank ordering about 90% of the time [19]. Suppose that the distance between instances i and j is ranked kth $(1 \le k \le n(n-1)/2)$ in the initial feature space compared with all other distances, and it receives a rank $r_k\,(1 \le r_k \le n(n-1)/2)$ when being projected onto a feature subspace. Denote the rank ordered list of the initial point pattern as the vector $\mathbf{v}_0 = (1,2,...,n(n-1)/2)$ and that of its lower-dimensional projection as the vector $\mathbf{v} = (r_1, r_2, ..., r_{n(n-1)/2})$. Using only these ranks, we construct a measure of the extent to which the projected point pattern falls short of a perfect match of the initial one. This normalized inter-angular measure is defined as

$$I_a = \frac{angle(\mathbf{v}_0, \mathbf{v})}{\underset{\mathbf{v}}{Max}\{angle(\mathbf{v}_0, \mathbf{v})\}}, \tag{4.17}$$

where $angle(\mathbf{v}_0, \mathbf{v})$ is the angle between $\mathbf{v}_0$ and $\mathbf{v}$. When $\mathbf{v}_0 = \mathbf{v}$, $I_a = 0$ and the projection is identical to the initial pattern as far as I_a can tell. When

* We assume that the orientation from one event to another is not important.

$\mathbf{v} = (n(n-1)/2, n(n-1)/2-1,...,1)$, then $I_a = 1$ and the projection disagrees with the initial pattern the most.

5. Density Estimation

In Section 3, we described a new framework for spatial-temporal prediction that extends knowledge discovery into feature space. We describe the estimation of individual model components in this section. Only key features picked out as the result of the feature selection step are used for model building. However, in this section we keep the same notation where no confusion arises in order to keep the amount of notation to minimum.

5.1. The "best" partition of the data

The estimators we consider can effectively accommodate local variations in the data. However, these estimators require that we estimate the appropriate number of distinct local (covariance) structures from the data first, unless we know them *a priori* (e.g., crime analysts may tell us how many groups of offenders are likely to be represented by the data). We use hierarchical clustering with a selection rule to achieve this.

0: Group n instances into n clusters $\Omega_1, \Omega_2, ..., \Omega_n$, each of which contains one instance.
1: Find the nearest pair of **distinct** clusters, say Ω_u and Ω_v, merge Ω_u and Ω_v and call the result Ω_u, delete Ω_v and decrement the number of clusters by one.
2: If the number of clusters equals one, then stop; otherwise, return to 1.

Figure 5.1. Basic operations of hierarchical clustering algorithms.

The basic operations of hierarchical clustering algorithms are similar, and are outlined in Figure 5.1 [14]. The difference between algorithms lies in the definition of

cluster-to-cluster distance (i.e., what we mean by "nearest"). For a data set of n instances, a hierarchical clustering algorithm generates a succession of n partitions $P_0, P_1, ..., P_{n-1}$, where $P_0, P_1, ..., P_{n-1}$ contain n, n-1, ..., 1 clusters, respectively. What we hope to find is the one partition that best represents the "natural structure" of the data set. In other words, we need a rule to select a partition from the complete hierarchy of n partitions, or better yet, stop merging clusters further as soon as we find the best partition. The "stopping rule" suggested by Mojena [29] is among the most satisfactory proposals. Let α_j ($j = 0,1,..,n-1$) denote the minimum distance between two clusters in partition P_j. α_j is called the *fusion level* at the stage with $n-j$ clusters. P_{j+1} is obtained by merging the two clusters in P_j distanced by α_j. In detail the stopping rule is to select the first partition P_j in the hierarchy satisfying

$$\alpha_{j+1} > \bar{\alpha} + k \cdot s_\alpha \qquad (5.1)$$

where α_{j+1} is the fusion level if further fusion were to take place, $\bar{\alpha}$ and s_α are, respectively, the mean and unbiased standard deviation of $\alpha_0, \alpha_1, ..., \alpha_{n-1}$ and k is a constant. The partition selected according to (5.1) consists of $n-j$ clusters. The constant k is usually set to 1.25, as recommended by Milligan and Cooper [28]. The rationale of Mojena's proposal is to look for significant "jump" in the α series.

Strictly speaking, Mojena's rule is not a *stopping* rule; but rather, it is *selection* rule. This is because $\bar{\alpha}$ and s_α in (5.1) are calculated based on the complete series $\alpha_0, \alpha_1, ..., \alpha_{n-1}$ which correspond to the complete hierarchy of n partitions. When n is

large, we find that using the partial α series to calculate the mean and standard deviation yields similar result. The revised Mojena's rule is then given as

$$\alpha_{j+1} > \overline{\alpha}_j + k \cdot s_{\alpha j} \tag{5.2}$$

where $\overline{\alpha}_j$ and $s_{\alpha j}$ are the mean and unbiased standard deviation of $\alpha_0, \alpha_1, ..., \alpha_j$. This is a bona fide stopping rule.

5.2. First order spatial transition density $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$

We consider two classes of models for estimating the first order spatial transition density. Both classes play an important role in modeling data that are believed to come from multiple underlying categories and sources. The first class is called *finite mixture distributions* (see [15], [27], [43]). These distributions are superpositions of (usually simpler) component distributions. A finite mixture probability density function (or mass function in the case of discrete sample space) has the form

$$f(\mathbf{x};\boldsymbol{\pi},\Theta) = \sum_{j=1}^{C} \pi_j f_j(\mathbf{x};\boldsymbol{\theta}_j) \tag{5.3}$$

where $\pi_j > 0$, $j = 1,2,...,C$, $\pi_1 + \pi_2 + ... + \pi_C = 1$, $\boldsymbol{\pi} = [\pi_1 \quad ... \quad \pi_C]'$, $\Theta = [\boldsymbol{\theta}_1 \quad ... \quad \boldsymbol{\theta}_C]$. $f_j(\mathbf{x};\boldsymbol{\theta}_j)$ is the *j*th *component density* with the set $\boldsymbol{\theta}_j$ of parameters and $\pi_1, \pi_2, ..., \pi_C$ are *mixing weights*. Θ is the collection of all *component parameters*. To fit a finite mixture distribution one needs to find the number C of component densities first. In our case this is done by applying hierarchical clustering to the data $\{\mathbf{x}_i : i = 1,2,..,n\}$.

More often than not, it is required that all component densities belong to the same parametric family. Suppose the vector $\mathbf{x}$ represents p numeric variables. The most widely used continuous finite mixture models are Gaussian mixture models (GMM), where the C

component densities are postulated as multivariate Gaussian distributions. In particular, the jth component density is

$$f_j(\mathbf{x};\boldsymbol{\theta}_j) = f_j(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_j|^{1/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)}, \; j = 1,2,...,C \quad (5.4)$$

where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Sigma}_j$ the covariance matrix. The parameter set $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$. Latent class models (LCM) (see [13]) are an important class of discrete finite mixture models. Suppose the vector $\mathbf{x}$ represents p categorical variables, and lth variable $[\mathbf{x}]_l$ (lth dimension of $\mathbf{x}$) takes on g_l distinct values $0,1,\ldots,g_l-1$. Assume that the variables are independent and the outcomes of each variable are also independent. Then $\mathbf{x}$ has a finite mixture distribution (5.3) with the jth component density being

$$f_j(\mathbf{x};\boldsymbol{\theta}_j) = f_j(\mathbf{x};\boldsymbol{\delta}_{j1},...,\boldsymbol{\delta}_{jp}) = \prod_{l=1}^{p}\prod_{k=0}^{g_l-1}\mathbf{1}_{\{[\mathbf{x}]_l=k\}}\delta_{jlk}, \quad j = 1,2,...,C \quad (5.5)$$

where $\boldsymbol{\delta}_{jl} = [\delta_{jl0} \;\; ... \;\; \delta_{jlg_l-1}]'$, $l = 1,2,...,p$, $j = 1,2,...,C$, and δ_{jlk}, $k = 0,1,...,g_l-1$, are the probabilities of $[\mathbf{x}]_l = k$. $\sum_{k=0}^{g_l-1}\delta_{jlk} = 1$. The indicator variable $\mathbf{1}_{\{[\mathbf{x}]_l=k\}}$ equals 1 when $[\mathbf{x}]_l = k$ and 0 otherwise. The parameter set $\boldsymbol{\theta}_j = \{\boldsymbol{\delta}_{j1},...,\boldsymbol{\delta}_{jp}\}$.

To estimate the parameters $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \;\; ... \;\; \boldsymbol{\theta}_C]$, we first calculate these quantities in accordance with the clusters, and then update them iteratively until the log likelihood $L = \sum_{i=1}^{n}\log f(\mathbf{x}_i;\boldsymbol{\pi},\boldsymbol{\Theta})$ converges to a stationary point. This numeric maximum likelihood method is known as EM algorithm [12]. We give the detail of the procedure in Figure 5.1. For the situation where mixed variable types are present, it is trivial to combine GMM and LCM provided that the numeric dimensions are independent of the categorical ones.

The second class of techniques that we use to estimate the first order spatial transition density belongs to the family of nonparametric models and is known as *filtered kernal estimators* (FKE) [25]. They take the general form

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}\frac{\rho_j(\mathbf{x}_i)}{|\mathbf{H}_j|}K\left(\mathbf{H}_j^{-1}(\mathbf{x}-\mathbf{x}_i)\right) \tag{5.6}$$

where $\mathbf{H}_j$, $j=1,2,...,C$, are C $p\times p$ nonsingular *local bandwidth matrices* and $\rho_j(\mathbf{x})$, $j=1,2,...,C$, which satisfy

$$0\le\rho_j(\mathbf{x})\le 1 \quad \text{and} \quad \sum_{j=1}^{C}\rho_j(\mathbf{x})=1 \tag{5.7}$$

for all $\mathbf{x}$, are *filtering functions*. Local bandwidth matrices contain posterior parameter settings that enforce localized smoothness for locally varied regions of the support of the

0: Let

$m = 0$,

$\pi_j^{(m)} = n_j / n$, $j = 1,2,...,C$,

(GMM) $\boldsymbol{\mu}_j^{(m)} = \left(1/n_j\right)\sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{x}_i \in \Omega_j\}} \mathbf{x}_i$, $j = 1,2,...,C$,

(GMM) $\boldsymbol{\Sigma}_j^{(m)} = \left(1/n_j\right)\sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{x}_i \in \Omega_j\}} \left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j\right)'$, $j = 1,2,...,C$,

(LCM) $\delta_{jlk}^{(m)} = \left(1/n_j\right)\sum_{i=1}^{n} \mathbf{1}_{\{\mathbf{x}_i \in \Omega_j\}} \mathbf{1}_{\{[\mathbf{x}_i]_l = k\}}$, $l = 1,2,..., p$, $j = 1,2,...,C$, $k = 0,1,..., g_l - 1$,

$f\left(\mathbf{x}_i; \boldsymbol{\pi}^{(m)}, \boldsymbol{\Theta}^{(m)}\right) = \sum_{j=1}^{C} \pi_j^{(m)} f_j\left(\mathbf{x}_i; \boldsymbol{\theta}_j^{(m)}\right)$,

$L^{(m)} = \sum_{i=1}^{n} \log f\left(\mathbf{x}_i; \boldsymbol{\pi}^{(m)}, \boldsymbol{\Theta}^{(m)}\right)$.

1: Let

$w_{ij}^{(m+1)} = \pi_j^{(m)} f_j\left(\mathbf{x}_i; \boldsymbol{\theta}_j^{(m)}\right) / f\left(\mathbf{x}_i; \boldsymbol{\pi}^{(m)}, \boldsymbol{\Theta}^{(m)}\right)$, $i = 1,2,...,n$, $j = 1,2,...,C$,

$n_j^{(m+1)} = \sum_{i=1}^{n} w_{ij}^{(m+1)}$, $j = 1,2,...,C$,

$\pi_j^{(m+1)} = n_j^{(m+1)} / n$, $j = 1,2,...,C$,

(GMM) $\boldsymbol{\mu}_j^{(m+1)} = \left(1/n_j^{(m+1)}\right)\sum_{i=1}^{n} w_{ij}^{(m+1)} \mathbf{x}_i$, $j = 1,2,...,C$,

(GMM) $\boldsymbol{\Sigma}_j^{(m+1)} = \left(1/n_j^{(m+1)}\right)\sum_{i=1}^{n} w_{ij}^{(m+1)} \left(\mathbf{x}_i - \boldsymbol{\mu}_j^{(m+1)}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}_j^{(m+1)}\right)'$, $j = 1,2,...,C$,

(LCM) $\delta_{jlk}^{(m+1)} = \left(1/n_j^{(m+1)}\right)\sum_{i=1}^{n} w_{ij}^{(m+1)} \mathbf{1}_{\{[\mathbf{x}_i]_l = k\}}$, $l = 1,2,..., p$, $j = 1,2,...,C$, $k = 0,1,..., g_l - 1$,

$f\left(\mathbf{x}_i; \boldsymbol{\pi}^{(m+1)}, \boldsymbol{\Theta}^{(m+1)}\right) = \sum_{j=1}^{C} \pi_j^{(m+1)} f_j\left(\mathbf{x}_i; \boldsymbol{\theta}_j^{(m+1)}\right)$,

$L^{(m+1)} = \sum_{i=1}^{n} \log f\left(\mathbf{x}_i; \boldsymbol{\pi}^{(m+1)}, \boldsymbol{\Theta}^{(m+1)}\right)$.

2: If $L^{(m+1)} - L^{(m)} < \varepsilon$ for some small $\varepsilon > 0$, stop; otherwise, $m = m + 1$ and return to 1.

Figure 5.1. EM algorithm for fitting GMM and LCM.

true density. The filtering functions can be interpreted as prior weights over variations of local smoothness. As a special case (when $\mathbf{H}_j = \text{diag}[h_{j1} \quad ... \quad h_{jp}]$, $j = 1,2,...,C$), the *filtered product kernel estimators* or FPK estimators are given as

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}\frac{\rho_j(\mathbf{x}_i)}{h_{j1}\cdots h_{jp}}\left\{\prod_{l=1}^{p}K\left(\frac{[\mathbf{x}]_l - [\mathbf{x}_i]_l}{h_{jl}}\right)\right\} \tag{5.8}$$

where h_{jl} ($j = 1,2,...,C, l = 1,2,..,p$) is a local bandwidth for the lth dimension $[\mathbf{x}]_l$ of the jth locally varied region. The underlying assumption for FPK estimators is that all dimensions are mutually independent. In principle FKE have advantages over *both standard kernel estimators* (SKE) as well as *variable kernel estimators* (VKE). SKE use a global bandwidth matrix and are not suitable for handling multi-modal and locally varied data. VKE require a distinct bandwidth matrix for each data point and it is not always clear how to best incorporate *a priori* information about local smoothness into these estimators

We only consider FPK estimators in this paper. Suppose the data $\{\mathbf{x}_i : i = 1,2,..,n\}$ have been partitioned into C clusters $\Omega_1, \Omega_2, ..., \Omega_C$. Let n_j be the number of instances in cluster Ω_j. We construct a FPK estimator using the procedure illustrated in Figure 5.2. Note that the procedure does not necessarily give an optimal estimator (under any assumption on the true density) in terms of asymptotic mean integrated squared error or AMISE. However, it should lead to a good representation of the true density because:

- Using either a finite mixture model or the indicator function to construct the filtering functions should capture a reasonable amount of local variations among the smoothing we need to impose on the locally varied regions.

- Consider the data in the cluster Ω_j alone. If we were to fit a Gaussian product kernel estimator to those data, (5.11) would give the optimal bandwidths in the AMISE sense assuming the true density is multivariate Gaussian (see [39]). These estimators should capture sufficient local smoothness for the FPK estimators.

1: Derive the filtering functions in either of the following two ways:

- Fit a finite mixture model to the data $g(\mathbf{x}) = \sum_{j=1}^{C} \pi_j g_j(\mathbf{x})$. Set

$$\rho_j(\mathbf{x}) = \pi_j g_j(\mathbf{x}) / g(\mathbf{x}), \quad j = 1,2,...,C. \tag{5.9}$$

- Let the indicator $\mathbf{1}_{\{\mathbf{x}\in\Omega_j\}}$ be 1 if $\{\mathbf{x} \in \Omega_j\}$ and 0 otherwise. Set

$$\rho_j(\mathbf{x}) = \mathbf{1}_{\{\mathbf{x}\in\Omega_j\}}, \quad j = 1,2,...,C. \tag{5.10}$$

We term this special FPK estimator *weighted product kernel estimator* or WPK estimator.

2. Estimate local bandwidths using local data in each cluster. To wit,

$$\hat{h}_{jl} = \left(\frac{4}{p+2}\right)^{1/(p+4)} \hat{\sigma}_{jl} n_j^{-1/(p+4)}, \; l = 1,2,...,p, \; j = 1,2,...,C. \tag{5.11}$$

where $\hat{\sigma}_{jl}$ is the standard deviation of the lth variable $[\mathbf{x}]_l$ using data $\{[\mathbf{x}_i]_l : \mathbf{x}_i \in \Omega_j, i = 1,2,...,n\}$.

Figure 5.2. Procedure for constructing a FPK estimator.

5.3. Spatial interaction probabilities $\Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\}$, $j = 1,2,...,C$

For either finite mixture or filtered kernel estimators, models of distinct local structures are readily available. Notice that each local structure corresponds to a clique in a

clustering point pattern. Spatial interaction probabilities are estimated from these "local" models.

When a finite mixture distribution is used to model first order transition density $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$, spatial interaction probabilities are given as

$$\Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\} = \pi_j f_j(\mathbf{x}_{n+1};\boldsymbol{\theta}_j)/f(\mathbf{x}_{n+1};\boldsymbol{\pi},\boldsymbol{\Theta}), \quad j = 1,2,...,C. \tag{5.12}$$

When a filtered kernel estimator is used to model first order transition density $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$, spatial interaction probabilities are given as

$$\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n) = \hat{f}_j(\mathbf{x}_{n+1})/\hat{f}(\mathbf{x}_{n+1}), \quad j = 1,2,...,C \tag{5.13}$$

where $\hat{f}_j(\mathbf{x}_{n+1})$, $j = 1,2,...,C$, are specified as follows, in correspondence to the generic form (5.6).

$$\hat{f}_j(\mathbf{x}_{n+1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_j(\mathbf{x}_i)}{|\mathbf{H}_j|} K(\mathbf{H}_j^{-1}(\mathbf{x}_{n+1} - \mathbf{x}_i)), \quad j = 1,2,...,C. \tag{5.14}$$

For filtered product kernel estimators (5.8),

$$\hat{f}_j(\mathbf{x}_{n+1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_j(\mathbf{x}_i)}{h_{j1}\cdots h_{jp}} \left\{\prod_{l=1}^{p} K\left(\frac{[\mathbf{x}_{n+1}]_l - [\mathbf{x}_i]_l}{h_{jl}}\right)\right\}, \quad j = 1,2,...,C. \tag{5.15}$$

5.4. Second order spatial transition densities $\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1})$, $j = 1,2,...,C$

Two models developed by Fiksel [16] can be used to estimate second order spatial transition densities $\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1})$, $j = 1,2,...,C$. Although other models are likely, Fiksel's models are among the simplest that incorporate the "journey to event" assumption we gave in Section 3. The instant-model also takes into account the assumption concerning "lingering period to resume act."

Remember from Section 3 that D_n and T_n, the sets of locations and times of n past events respectively, are partitioned into C disjoint subsets $\{D_n^{(j)} : j = 1,2,...,C\}$ and $\{T_n^{(j)} : j = 1,2,...,C\}$ in correspondence to the clustering pattern $\{\chi_n^{(j)} : j = 1,2,...,C\}$ in feature space. We need to establish C second order spatial transition density models, one based on each pair of data subsets $D_n^{(j)}$ and $T_n^{(j)}$. To simplify the notation we will drop the subset label j from individual data units. This should not be confusing as long as the reader bears in mind that the presentation below is applicable to any pair of data subsets $D_n^{(j)}$ and $T_n^{(j)}$.

Let the number of data units in $D_n^{(j)}$ and $T_n^{(j)}$ be m and the locations and times of past events contained in these subsets be $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m$ and $t_1, t_2, ..., t_m$, where $t_1 < t_2 < ... < t_m$ and $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m$ are ordered by $t_1, t_2, ..., t_m$. According to Fiksel's order-model, we postulate the following function for the second-order spatial transition density

$$\psi_n^{(12)}\left(\mathbf{s} \middle| D_n^{(j)}, T_n^{(j)}, t\right) = \varphi_m\left(\mathbf{s} \middle| \mathbf{s}_1, ..., \mathbf{s}_m\right) = \frac{\lambda^2}{2\pi m} \sum_{i=1}^{m} e^{-\lambda \|\mathbf{s} - \mathbf{s}_i\|} \tag{5.16}$$

where $t > t_m$ is a future event's time of occurrence and $\|\mathbf{s} - \mathbf{s}_i\|$ the distance from that future event's location $\mathbf{s}$ to an older event location $\mathbf{s}_i$ ($i = 1,2,...,m$). Notice that if longitudes and latitudes are used to designate locations, they need to be transformed into UTM (Universal Transverse Mercator) coordinates before a distance measure can be calculated. This is called order-model since only the temporal order of the events is considered. The likelihood of observing the previous m events is then given as

$$L(\mathbf{s}_1, ..., \mathbf{s}_m; \lambda) = \varphi_{m-1}\left(\mathbf{s}_m \middle| \mathbf{s}_1, ..., \mathbf{s}_{m-1}\right) \cdots \varphi_1\left(\mathbf{s}_2 \middle| \mathbf{s}_1\right) = \prod_{i=1}^{m-1} \frac{\lambda^2}{2\pi i} \sum_{k=1}^{i} e^{-\lambda d_{i+1,k}} \tag{5.17}$$

where $d_{i+1k} = \|\mathbf{s}_{i+1} - \mathbf{s}_k\|$. The maximum likelihood estimate $\hat{\lambda}$ of the parameter λ is obtained by maximizing (5.17). Differentiating (5.17) with respect to λ and equating the result to zero yields

$$\hat{\lambda} = 2(m-1)\left\{\sum_{i=1}^{m-1} \frac{\sum_{k=1}^{i} d_{i+1k} e^{-\hat{\lambda} d_{i+1k}}}{\sum_{k=1}^{i} e^{-\hat{\lambda} d_{i+1k}}}\right\}^{-1} \tag{5.18}$$

which can be solved numerically by fix-point iteration (see [4]).

A second model in [16] is known as instant-model and it utilizes the actual values of the series $t_1, t_2, ..., t_m$. Based on this model, we postulate that the second order spatial transition density takes on the form

$$\psi_n^{(12)}\left(\mathbf{s}|D_n^{(j)}, T_n^{(j)}, t\right) = \eta_m\left(\mathbf{s}|\mathbf{s}_1, ..., \mathbf{s}_m, t_1, ..., t_m, t\right) = \frac{\lambda^2}{2\pi \sum_{i=1}^{m} e^{-\tau(t-t_i)}} \sum_{i=1}^{m} e^{-\lambda\|\mathbf{s}-\mathbf{s}_i\| - \tau(t-t_i)}. \tag{5.19}$$

Similarly, the likelihood function is given as

$$\begin{aligned} L(\mathbf{s}_1, ..., \mathbf{s}_m, t_1, ..., t_m; \lambda, \tau) &= \eta_{m-1}\left(\mathbf{s}_m|\mathbf{s}_1, ..., \mathbf{s}_{m-1}, t_1, ..., t_{m-1}, t_m\right) \cdots \eta_1\left(\mathbf{s}_2|\mathbf{s}_1, t_1, t_2\right) \\ &= \prod_{i=1}^{m-1} \frac{\lambda^2}{2\pi \sum_{k=1}^{i} e^{-\tau \Delta_{i+1k}}} \sum_{k=1}^{i} e^{-\lambda d_{i+1k} - \tau \Delta_{i+1k}} \end{aligned} \tag{5.20}$$

where $\Delta_{i+1k} = t_{i+1} - t_k$. This is to be maximized to yield the maximum likelihood estimates $\hat{\lambda}$ and $\hat{\tau}$.

5.5. Temporal transition density $\psi_n^{(2)}\left(t_{n+1}|T_n\right)$

Under appropriate assumptions, both stochastic process models (e.g., Poisson) and time series techniques (e.g., first-order autoregressive, ARIMA) may be applied. Notice that the temporal transition density only depends on past event times and thus is constant for all locations within the study region (Assumption 3.2). When only the relative

magnitudes of spatiotemporal transition densities are important, temporal transition element may be omitted.

5.6. Geographic-space feature density $\kappa_n(\mathbf{x}_{n+1})$

In general, estimation of this density needs sampling over the study region. For example, we may obtain feature values for the locations on a regular grid over the study region. We may then fit a density function to these sample values using either finite mixture or filtered kernel method.

6. Conclusion

In this paper, the problem of predicting the likelihood of space-time random events has been considered. Unlike the traditional approach that overlooks event-related features, the current work has aimed at bringing feature-space analysis into space-time prediction. This added dimension ensures that event initiation patterns frequently hidden in feature data can be discovered and used to inform future event occurrences. The new approach is able to identify potential event locations far away from the past event locations. To demonstrate this claim, we have implemented a version of the model and tested it on a simulated data set of criminal incidents in the Charlottesville-Albemarle region of central Virginia. The result was reported in [24]. A real-world application and evaluation of our model in the regional crime analysis domain will be reported in another paper.

This work represents a first step in the introduction of feature space analysis into space-time event prediction. There is ample room for future research. The following several directions warrant further investigation.

- Hybrid measures for dimensionality reduction of unsupervised data hold promise for generating the suitable number of features to use. By hybrid measures, we mean those measures generated by combining a measure of cohesiveness (of a point pattern) and a measure of disagreement (between two point patterns). Preliminary exploration of this area can be examined in [23]. Further work is desirable to find a sound weighting scheme or alternative hybrid forms for this category of measures.
- In the current application of filtered kernel density estimators, local bandwidths are obtained by applying Gaussian bandwidth selection rule on local data (see (5.11)). Although this is a good heuristic, it would be nice if an optimal bandwidth rule could be found for this class of density estimators based on resampling criteria (e.g., bootstrap, cross validation).
- The second-order spatial transition densities in our model are estimated by Fiksel's spatial transition models. Alternative models (e.g., from Regional Economics literature) may be used provided that the parameter estimation problem can be easily solved.
- A previously developed case matching methodology [18] may be integrated into the prediction model. In particular, the current model puts equal weight on every feature. The case matching methodology may be used to find an appropriate weight for each feature in order to bring partitions in feature space closer to reality (i.e., the actual preferences of event initiators).
- The current model copes with temporal heterogeneity in an indirect fashion. First, the gross effect of all temporal features may be incorporated collectively into a temporal transition density model. Second, assuming that the spatial feature values are

dynamically changing with time and the choice of an event location is dependent on the current feature values of that location, temporal heterogeneity does play a role in event site selection decisions through a non-static study region. We would like to test our model in real-world settings where static or non-static study regions are involved and see how well it performs.

- Again, unless we relax the assumption that temporal evolution is independent of geographic locations (Assumption 3.1), the temporal transition density component remains constant for every location in the study region. It changes over time and subsequently changes the magnitude of the spatial-temporal transition density for every location by a common factor. To explicitly consider temporal heterogeneity, temporal features need to be identified, selected, and analyzed. Although it may not look so obvious, it seems that the feature space analysis that we presented for spatial features will follow through for temporal features. Further efforts in this direction should be rewarding.

References

[1] B. Abraham, "The exact likelihood function for a space-time model," *Metrika*, vol. 30, pp. 239-243, 1983.

[2] M. Amir, *Patterns in Forcible Rape*. Chicago, IL: University of Chicago Press, 1971.

[3] L. A. Aroian, "Time series in m dimensions: Definitions, problems, and prospects," *Communications in Statistics, Simulation and Computation*, vol. B9, pp. 453-465, 1980.

[4] N. S. Asaithambi. *Numerical Analysis: Theory and Practice*. Fort Worth, TX: Harcourt College Publishers, 1995.

[5] J. Baldwin and A. Bottoms, *The Urban Criminal: A Study in Sheffield*. London: Tavistock Publications, 1976.

[6] M. Ben-Bassat and J. Raviv, "Renyi's entropy and the probability of error," *IEEE Transactions on Information Theory*, vol. 24, pp. 324-331, 1978.

[7] R. Birkenhead, "Similarity between categorizations," in *Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics*, 1998, pp. 4005-4009.

[8] P. Brantingham and P. Brantingham, "Spatial patterns of burglary," *Howard Journal of Penology and Crime Prevention*, vol. 14, pp. 11-24, 1975.

[9] D. Capone and W. Nichols, "Urban structure and criminal mobility," *American Behavioral Scientist*, vol. 20, pp. 199-213, 1976.

[10] A. D. Cliff, P. Hagget, J. K. Ord, K. A. Bassett, and R. B. Davies, *Elements of Spatial Stucture: A Quantitative Approach*. Cambridge, U.K.: Cambridge University Press, 1975.

[11] M. Dash, H. Liu, and J. Yao, "Dimensionality reduction of unsupervised data," in *Proceedings of Ninth IEEE International Conference on Tools with Artificial Intelligence*, 1997, pp. 532-539.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)," *Journal of Royal Statistical Society, Series B*, vol. 39, pp. 1-38, 1977.

[13] B. S. Everitt, *An Introduction to Latent Variable Models*. London: Chapman and Hall, 1984.

[14] B. S. Everitt, *Cluster Analysis*, 3rd ed. London: Edward Arnold, 1991.

[15] B. S. Everitt and D. J. Hand, *Finite Mixture Distributions*. London: Chapman and Hall, 1981.

[16] T. Fiksel, "Simple spatial-temporal models for sequences of geological events," *Elektronische Informationsverarbeitung und Kybernetik*, vol. 20, pp. 480-487, 1984.

[17] H. P. Friedman and J. Rubin, "On some invariant criteria for grouping data," *Journal of American Statistical Association*, vol. 62, pp. 1159-1178, 1967.

[18] S. C. Hagen, "*A suspect matching tool for robbery data*," Master's thesis, Department of Systems Engineering, University of Virginia, Charlottesville, VA, 1999.

[19] J. G. Kreifeldt, S. H. Levine, K. Nah, and L. Liu, "Determining the similarity of point patterns using internal, nonmetric representations: Preliminary results," in

Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics, 1998, pp. 4492-4497.

[20] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Pyschometrika*, vol. 29, pp. 1-27, 1964.

[21] J. B. Kruskal, "Non-metric multidimensional scaling: A numerical method," *Pyschometrika*, vol. 29, pp. 115-129, 1964.

[22] J. L. LeBeau, "The journey to rape: Geographic distance and the rapist's methods of approaching the victim," *Journal of Police Science and Administration*, vol. 15, pp. 129-136, 1987.

[23] H. Liu, "*Space-time point process modeling: Feature selection and transition density Estimation*," Ph.D. dissertation, Department of Systems Engineering, University of Virginia, Charlottesville, VA, 1999.

[24] H. Liu and D. E. Brown, "Spatial-temporal event prediction: A new model," in *Proceedings of the 1998 IEEE International Conference on Systems, Man and Cybernetics*, 1998, pp. 2933-2937.

[25] D. J. Marchette, C. E. Priebe, G. W. Rogers, and J. L. Solka, "Filtered kernel density estimation," *Computational Statistics*, vol. 11, pp. 95-112, 1996.

[26] R. J. Martin and J. E. Oeppen, "The identification of regional forecasting models using space-time correlation functions," *Transactions of the Institute of British Geographers*, vol. 66, pp. 95-118, 1975.

[27] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, Inc., 1988.

[28] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, pp. 159-179, 1985.

[29] R. Mojena, "Hierarchical grouping methods and stopping rules: An evaluation," *Computer Journal*, vol. 20, pp. 359-363, 1977.

[30] T. Molumby, "Patterns of crime in a university housing project," *American Behavioral Scientist*, vol. 20, pp. 247-259, 1976.

[31] O. Newman, *Defensible Space: Crime Prevention through Urban Design*. New York: Macmillan, 1972.

[32] P. E. Pfeifer and S. J. Deutsch, "Identification and interpretation of first order space-time ARMA models," *Technometrics*, vol. 22, pp. 397-408, 1980.

[33] P. E. Pfeifer and S. J. Deutsch, "A three-stage iterative procedure for space-time modelling," *Technometrics*, vol. 22, pp. 35-47, 1980.

[34] A. Renyi, "On measures of entropy and information," in *Proceeding of 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1960, pp. 547-561.

[35] T. A. Repetto, *Residential Crime*. Cambridge, MA: Ballinger, 1974.

[36] D. K. Rossmo, "Target patterns of serial murders: A methodological model," *American Journal of Criminal Justice*, vol. 17, no. 2, pp. 1-21, 1993.

[37] D. K. Rossmo, "Targeting victims: Serial killers and the urban environment," in *Serial and Mass Murder: Theory, Research, and Policy*, T. O'Reilly-Flemming and S. Egger, Eds. Toronto: University of Toronto Press, 1994.

[38] H. A. Scarr, *Patterns in Burglary*, 2nd ed. Washington, D.C.: U.S. Department of Justice, 1973.

[39] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley, 1992.

[40] R. N. Shepard, "Multidimensional scaling, tree-fitting, and clustering. *Science*," vol. 210, pp. 390-398, 1980.

[41] D. S. Stoffer, "Maximum likelihood fitting of STARMAX models to incomplete space-time series data," in *Time Series Analysis: Theory and Practice 6*, O. D. Anderson, J. K. Ord, and E. A. Robinson, Eds. Amsterdam: North-Holland, 1985, pp. 283-296.

[42] D. S. Stoffer, "Estimation and identification of space-time ARMAX models in the presence of missing data," *Journal of the American Statistical Association*, vol. 81, pp. 762-772, 1986.

[43] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley, 1985.

Criminal Incident Prediction Using Preference Discovery

Hua Liu
Lucent Technologies, Inc.
One Main Street
Cambridge, MA 02142
hualiu@lucent.com

Donald E. Brown
Department of Systems Engineering
University of Virginia
Charlottesville, VA 22903
brown@virginia.edu

Abstract: Criminal incidents are human-initiated events that may be assumed to be dictated by certain preferences over space and time. In this paper we first establish the correspondence between a set of preferences and a cluster of values of certain key event features and then present a point process transition density model for space-time event prediction that hinges upon preference discovery or the point pattern in the key feature space. The added dimension of feature space analysis enables our model to outperform the traditional "hot spots" approach, as demonstrated statistically by the real-world application involving breaking and entering incidents in Richmond, Virginia. Being able to accommodate all measurable features, identify the key features and quantify their relationship with criminal incident occurrence over space and time, our model provides the basis for developing and testing theories of criminal activity.

1. Introduction

Law enforcement agencies have increasingly acquired database management systems (DBMS) and geographic information systems (GIS) to support their crime analytic capabilities. These agencies use such systems to monitor current crime activity and develop collaborative strategies with local communities for combating crime. However, in general these strategies tend to be reactive rather than proactive. A more proactive approach requires early warning of trouble with sufficient lead-time to formulate a plan. Early warning, in turn, necessitates the development of predictive models in space and time that can inform law enforcement of pending "hot spots" and areas with declining crime activity.

Criminal incidents, like many other human-initiated events, are frequently linked with the preferences that event initiators (i.e., offenders) have for specific sites and

specific time slots in terms of certain spatial and temporal attributes (or features[1]) of those sites and time slots, respectively. The spatial aspect of this phenomenon has been well documented in criminology literature and supported by various spatial theories of criminal activity. One of the most complete discussions of spatial patterning in crime is contained in Brantingham and Brantingham (1984). From the standpoint of this work we are particularly interested in what they call the microspatial component of crime or the choice of crime locations by individual criminals. A number of researchers have documented and formulated descriptions for spatial decision making by criminals (see, for example, Brantingham and Brantingham, 1975; Molumby, 1976; Newman, 1972; Repetto, 1974; Scarr, 1973). Some have looked specifically at the question of distance from home to crime location (for example, Amir, 1971; Baldwin and Bottoms, 1976; Capone and Nichols, 1976; LeBeau, 1987; Rossmo, 1993; Rossmo, 1994). Taken together this impressive body of research shows that "target selection is a spatial information processing phenomenon." (Brantingham and Brantingham 1984, p.344). Essentially offenders have certain preferences in their site selection. These preferences can be defined in relation to a selected set of spatial attributes or features.

It is rather safe to say that offenders' preferences constitute an important piece of information to inform future site selection decisions by criminals. It is desirable that predictive models for crime incidents take advantage of this preference information, more specifically, the pattern revealed by the feature data of the observed incidents. Predictive models that fail to look into the feature data to address incident initiation preferences are inevitably not as intuitive and, quite possibly, do not predict as well as what we expect.

[1] We use the term features as a synonym for terms such as predictor or independent variables, which are commonly used in regression and linear modeling.

Such models are essentially variants of traditional pin-mapping techniques. They ignore feature data and basically map out the locations of past incidents and their vicinities as predicted criminal "hot spots," based on certain assumptions on spatial dependence. In this paper, we describe a space-time prediction model that we recently developed based on the theory of point patterns and multivariate density estimation. The model itself and the formal analysis that we propose for building the model establish an approach for discovering and representing criminal preferences as the functional relationships between demographic, economic, social, victim, and spatial variables and numerous measures of criminal activity. Our intent is not only to give a new statistical model that integrates feature space analysis into space-time prediction, but also to provide the critical infrastructure for building and testing the theories of criminal activity that compete with one another for use in the major law enforcement strategies. Therefore our model must be understandable to users, accurate, and testable with a variety of theoretical constructs from current research on crime. By understandable to users, we mean that the model cannot be a "black box." The user needs to understand how the inputs to the model are used to formulate a prediction. This is particularly important for testing theories of criminal activity. Our model must allow a variety of different features and our approach must include a way to test the effectiveness of these features. For example, our model should conveniently allow us to test a theory that says that proximity to bars or nightclubs contributes to crime of certain type in an area.

The remainder of this paper is organized as follows: In the next section, we take a closer look at the distributions of criminal incidents in temporal, geographic, and feature spaces, respectively, and explain intuitively how we may capture the incident initiation

preferences in feature space. In Section 3 we give a formal account of the criminal incident prediction problem and describe the assumptions and technical details of our model for solving the problem. In Section 4 we present a real-world application of our proposed model and the evaluation and comparison of our model against the traditional "hot spot" approach. Section 5 summarizes our modeling approach and the contributions of this approach to law enforcement and to solving space-time prediction problems in other domains.

2. Preference Discovery in Feature Space

Criminal incident prediction is usually carried out within a specified geographic region (e.g., a jurisdiction) and within a specified time range (e.g., a month) for a specified crime type. In practice, these boundary conditions are defined by law enforcement agencies. We term the geographic region of interest a *study region* or *geographic space* $D \subset \Re^2$, and the time range a *study horizon* $T \subset \Re^+$. To formally capture the criminal incident prediction problem, we regard the locations and times of the incidents of a specific type as vectors $(\mathbf{s}_1,t_1)$, $(\mathbf{s}_2,t_2)$, ... , $t_0 = 0 < t_1 < t_2 < \ldots$, where $\mathbf{s}_i \in D$ is the two dimensional location of incident i and t_i is the time of this incident. The incidents also have corresponding features (or marks) $\mathbf{x}_1$, $\mathbf{x}_2$, ... that describe the attributes of the incidents (e.g. distance to a road, type of residential community, etc.). Suppose that initially we have p measurable features $f_1, f_2, \ldots, f_p$ that are known or believed to be relevant to the occurrence of the incidents. Then the hyperspace formed by these p features is a (p-dimensional) *feature space* $\chi \subset \Re^p$. A subset of the initial feature set defines a *feature subspace*. Mathematically, taken together the locations, times, and

features of all incidents constitute a realization of a *marked space-time shock point process*.

We have mentioned in the introductory section that for many human-initiated events, one primary behavioral assumption is that *event initiators (e.g., offenders in crime scenario) choose the site and time of an event based upon a set of preferences over the values of the attributes (features) at alternative sites and times*. While the events of a marked space-time point process may be presented in three hyperspaces (time axis, geographic space and feature space), event initiation preferences are measured in feature space. Suppose that the initial set of features contains those attributes that the event initiators **actually** factor into their decision making. A set of preferences pertaining to a group of event initiators is defined when the subset of features actually considered by the group of event initiators and a partial ordering of available values for these features are specified. For a specific group of event initiators, if we knew their set of preferences (i.e., the subset of features and the partial order for the feature subset), we would examine all location-time combinations for their feature values and score them accordingly. However, without its knowledge, we must "discover" it from the data, more specifically, from the point pattern in feature space.

Preference discovery in feature space prompts two questions. First, which features are actually considered by a group of event initiators? We are never going to know with certainty the answer to this question. Our objective instead is to find the smallest feature subset (of the initial feature set) that is necessary and sufficient to account for the underlying pattern of event occurrences. This is known as *feature selection.* We term the selected feature subset the *key feature set* and the feature subspace defined by the key

feature set the *key feature space.* By definition, the underlying pattern of event occurrences should manifest itself most clearly in the key feature space. This leads to the second question: What kind of point pattern do we expect to see in the key feature space? The answer to this second question provides the basis for specifying a partial order for the key feature set. We give the formal models for the partial order in the next section.

To answer the second question, we make the following two assumptions: (1) *If multiple groups of event initiators are present, they make site selection decisions based on common set of features*, and (2) *preferences remain stable (stationary in probabilistic sense) over the study region and study horizon for each group of event initiators.* The first assumption is inevitable if we want to deal with multiple groups simultaneously. With the second "stationarity" assumption, we may conclude that given the data of repeated event initiation decisions by a group, the set of preferences of this specific group (or the underlying pattern of event occurrences) must manifest itself as a small-variation distribution of values in the key feature space. This small-variation distribution can be described as a *clique* in point process theory (or less formally as a *cluster*). If multiple groups with distinct preferences are present over the study region and study horizon, we expect to see a clustering (point) pattern with multiple cliques in the key feature space.

We illustrate the above observation in Figure 1, where we have assumed that initial feature set is the key feature set. Although the distribution of events on time axis as well as that in geographic space could very much lack any systematic pattern, stable and distinct clustering patterns should be observed in feature space. Each clique in feature space corresponds to a set of preferences. It is often the case that locations in close geographic proximity have similar feature values. Then neighbors in geographic space

are neighbors in feature space (e.g., s_6 and s_7). However, proximity in feature space does not necessarily translate into proximity in the geographic space (e.g., s_2 and s_5). It is quite possible that two locations that are far apart have the same feature values and thus it is only reasonable to assign an equal score to both locations if we extrapolate event occurrence based solely upon site selection preferences. The merit of integrating feature space information into space-time event prediction is that **potential** event areas (e.g., areas not previously struck as frequently by crimes but at high risk nevertheless) can be picked out. The same rationale applies to the analysis of event occurrences in time.

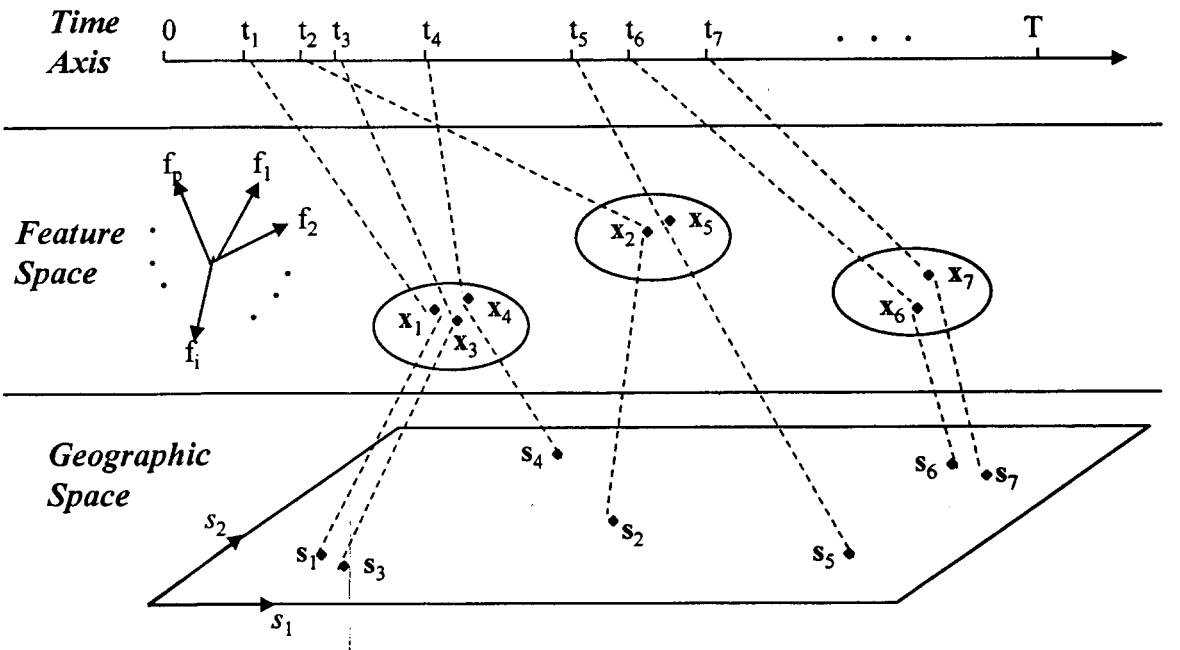

Figure 1. Event occurrences in three hyperspaces.

We have intentionally kept the subject of the discussion in this section as "(human-initiated) events" since preference discovery is not confined to the crime incident prediction scenario. We will give some other examples in the final section.

3. The Model

Criminal incidents (and other human-initiated events in a more general context) are random events in space and time. The quantity of general interest is naturally the likelihood that a future incident occurs within a study region and a study horizon, given the times, locations, and feature values of past incidents of the same type bounded by the same region and time range. Formally, this likelihood is the transition density of the marked space-time shock point process we mentioned earlier. Let $T_n = \{t_1, t_2, ..., t_n\}$, $D_n = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ and $\chi_n = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ where $\mathbf{s}_i = (s_{i1}, s_{i2})$ and $\mathbf{x}_i = [x_{i1} \quad ... \quad x_{ip}]'$. The transition density is defined as follows.

$$\psi_n(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n, \chi_n) \equiv \lim_{\nu(d\mathbf{s}_{n+1}), dt_{n+1} \to 0} \frac{\Pr\{N(d\mathbf{s}_{n+1}, dt_{n+1}) = 1 | D_n, T_n, \chi_n\}}{\nu(d\mathbf{s}_{n+1}) dt_{n+1}} \qquad (1)$$

where $\mathbf{s}_{n+1}$ and t_{n+1} are the location and the time of the next incident, respectively, $\nu(d\mathbf{s}_{n+1})$ is the Lebesgue measure of $d\mathbf{s}_{n+1}$ and $N(d\mathbf{s}_{n+1}, dt_{n+1})$ counts the incidents that happen within the infinitesimal region $d\mathbf{s}_{n+1}$ and the infinitesimal time interval dt_{n+1}. It is the probability that a single future incident occurs within specified infinitesimal region and specified infinitesimal time interval. In theory, "single" or uniquely identifiable events are ensured if we postulate a *simple* point process.

The discussion in this section focuses on two topics surrounding the transition density defined in (1). First, we give a model of the transition density. Such a model can be used to dynamically generate density estimates over space and time for the occurrence of future incidents. Second, we present criteria for evaluating and identifying which of the features have the most predictive or explanatory power. These two topics are closely related. From the empirical model building point of view, the second topic, known as

feature selection, is a preliminary step to the first topic and it determines which features or predictor variables should go into the transition density model. The model from the first topic, once built with certain features and tested on new data sets, could in turn be used to assess these features' contribution to prediction quality and justify their choice. In other words, the model formally specifies a partial order over the values of the selected features. It is important to note that for both topics we develop our models or techniques all in accordance with the notion of preference discovery that we illustrated in the last section.

3.1. The transition density model

The development of our model involves a multi-step componentization of the transition density (1) and the estimation of individual model components. This subsection describes the componentization and the next section deals with density estimation models for the components. We give both intuitive and formal descriptions of the process in the sequel.

The first step in the process is to separate spatial and temporal transitions. We postulate that the occurrences of criminal incidents over time and space are separable in the sense that

$$\psi_n\left(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n, \chi_n\right) = \psi_n^{(1)}\left(\mathbf{s}_{n+1} | D_n, \chi_n, T_n, t_{n+1}\right) \cdot \psi_n^{(2)}\left(t_{n+1} | T_n\right) \qquad (2)$$

where $\psi_n^{(1)}\left(\mathbf{s}_{n+1} | D_n, \chi_n, T_n, t_{n+1}\right)$ will be called *spatial transition density* and $\psi_n^{(2)}\left(t_{n+1} | T_n\right)$ *temporal transition density*. Equation (2) would be a standard Bayesian decomposition if the second term on the right-hand side were $\psi_n^{(2)}\left(t_{n+1} | D_n, \chi_n, T_n\right)$. D_n and χ_n were left out under two assumptions: (1) *The initial set of features does not contain any (inherently) temporal features*, and (2) *temporal evolution (transition) of the marked*

space-time shock point process does not depend on spatial (locational) evolution (transition). By "(inherently) temporal features," we mean features that "label" time intervals so that categorization of time instants can be obtained. Some examples are "seasons of the year," "weekdays / weekends," "segments of a day (e.g., morning / afternoon / night)." Technically, we may exclude all temporal features since the synthesized effect of different temporal categories on incident occurrence is contained in the time series $t_1, t_2, ..., t_n$ and the heterogeneity in the series may be incorporated by using ARIMA-like models to estimate temporal transition. The practical reason for not explicitly considering temporal features is that spatial component is more evident in the crime scenario and the need for validating spatial theories of criminal activity is more imminent. With temporal features excluded, χ_n becomes the collection of feature data that contains "site selection" preferences. Ignoring the temporal heterogeneity pertaining to a temporal feature also requires that this heterogeneity does not make site selection preferences unstable within the study horizon (i.e., does not render the "stationarity" assumption in Section 2 invalid). Otherwise, we must reduce study horizon to what is defined by a single category of the temporal feature (which may be considered a stationary time interval) and trim down the available data accordingly in order for our subsequent analysis to be tenable. The second assumption mentioned essentially says that spatial dependence arises from the integration of causal factors over time, but not vice versa. In the crime analysis scenario, this means that we do not regard the past crime intensity at a site as a direct factor to influence how soon criminals are going to strike again. However, this past behavior does tell us about the preferences of site selectors and

we directly model these preferences in the subsequent steps of the componentization below.

The second step of the componentization is concerned with how to model the spatial transition density $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n,\chi_n,T_n,t_{n+1})$. Intuitively speaking, our modeling philosophy is to use past site selection behavior to inform where events are likely to occur again. For the moment, assume that the features we select initially are the key features. By doing so, we postpone the feature selection task until next subsection. In the last section we have concluded that we expect to see a distinctive clustering pattern in the key feature space with each clique or cluster defines a set of event initiation preferences. Suppose that the set χ_n of feature vectors is partitioned into C disjoint subsets $\{\chi_n^{(j)} : j = 1,2,...,C$, each of which is mapped onto a clique in key feature space. Corresponding to $\{\chi_n^{(j)} : j = 1,2,...,C$, the set D_n (T_n) of locations (times) of past events is also partitioned into C disjoint subsets $\{D_n^{(j)} : j = 1,2,...,C$ ($\{T_n^{(j)} : j = 1,2,...,C\}$). Let $\mathbf{x}_{n+1}$ be the estimated feature vector at location $\mathbf{s}_{n+1}$ and instant t_{n+1}. Conditional on $\mathbf{x}_{n+1}$, the spatial transition density is assumed to take the form

$$\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n,\chi_n,T_n,t_{n+1}) = \alpha \cdot \psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n) \cdot \sum_{j=1}^{C} \psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)},T_n^{(j)},t_{n+1}) \Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)} \quad (3)$$

where $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$ is termed the *first order spatial transition density*[2] and reflects event intensity (i.e., first order effects) at $\mathbf{x}_{n+1}$ in feature space. $\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)},T_n^{(j)},t_{n+1})$, $j = 1,2,...,C$, are termed *second order spatial transition densities*, which reflect

[2] This is a probability mass function in the case of a discrete feature space. We shall use the term "density" in both continuous and discrete cases.

interaction (i.e., second order effects) of new event location $\mathbf{s}_{n+1}$ with past event locations in each $D_n^{(j)}$, respectively. $\Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}, \ j = 1,2,...,C,$ are *spatial interaction probabilities* or the probabilities that $\mathbf{x}_{n+1}$ and each $\chi_n^{(j)}$ form a clique in the feature space. α is a normalizing constant.

Model (3) incorporates all elements of site selection behavior and puts them into a formal framework — spatial point process theory. A spatial point pattern can be regarded as the result of first order effects coupled with second order effects. We model first order effects as the event initiators' site selection preferences or alternative sites' potential to attract future events (feature space analysis) rather than the average number of events already accumulated at alternative sites (geographic space analysis). This notion of site selection preferences is more fitting for prediction given that the same sets of preferences will carry on to t_{n+1} over the study region (see the "stationarity" assumption in the last section). We do not consider second order effects in feature space because we assume that *the spatial point process in the key feature space is Markovian over a small range*. Roughly speaking, this assumption ensures that in the key feature space, there are no second order effects (i.e., dependence or interaction) between cliques, and since the range (or clique radius) is small, only first order effects are important within each clique. This assumption formally characterizes the point pattern in the key feature space (or the site selection behavior revealed by feature space analysis).

The second order effects are modeled in geographic space. Notice that it is only appropriate to examine spatial dependence for events in the same feature-space clique (i.e., events initiated by the same group of people). However, due to the uncertainty associated with assigning a new event to a specific clique (or claiming that a specific

group is responsible for a new event), we weigh second order effects pertaining to individual cliques by the probabilities that quantify this uncertainty (i.e., spatial interaction probabilities). Technically, we estimate the weighted average of the second order effects of C thinned point processes in geographic space. A realization of each thinned point process is the set $D_n^{(j)}$ of events corresponding to those that form the clique $\chi_n^{(j)}$ in feature space.

The spatial transition density model (3) needs "prior" adjustment when the predicted feature values ($\mathbf{x}_{n+1}$'s) for all locations within the study region (D) do not form a uniform distribution. Let $\kappa_n(\mathbf{x}_{n+1})$ denote the probability density function of $\mathbf{x}_{n+1}$ over all predicted feature values for locations $\mathbf{s}_{n+1} \in D$. Non-uniformity of $\kappa_n(\mathbf{x}_{n+1})$ indicates certain feature values are more typical than others in the study region. Individual locations with typical feature values, if preferred by event initiators, should be at lower risk compared with those with rare feature values simply because event initiators have more choices over the region but they may engage themselves at only one location at any instant[3]. To put all locations on an equal footing, we adjust (3) as follows.

$$\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n,\chi_n,T_n,t_{n+1}) = \beta\cdot(1/\kappa_n(\mathbf{x}_{n+1}))\cdot\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$$
$$\cdot\sum_{j=1}^{C}\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)},T_n^{(j)},t_{n+1})\Pr\{\mathbf{x}_{n+1}\in\chi_n^{(j)}\} \qquad (4)$$

where β is a normalizing constant. When $\kappa_n(\mathbf{x}_{n+1})$ is uniform, (4) reduces to (3). $\kappa_n(\mathbf{x}_{n+1})$ can be easily estimated if all features are static over the study horizon. We use (3) when we do not have knowledge of $\kappa_n(\mathbf{x}_{n+1})$. We term $\kappa_n(\mathbf{x}_{n+1})$ the *geographic-space feature density.*

3.2. Density estimation

The equations (2), (3) and (4) collectively define our transition density model — a new framework for spatial-temporal event prediction that takes advantage of preference discovery in feature space. For our purpose, the estimation of the individual components involves the following four tasks:

(1) In the key feature space, partition the data into the "best" number (C) of clusters.

(2) Estimate the first order spatial transition density and the spatial interaction probabilities in the key feature space.

(3) Estimate the second order spatial transition densities in the geographic space.

(4) Estimate the geographic-space feature density where appropriate and feasible.

The astute reader may ask why we do not need to estimate temporal transition density. The answer is that generally we do need to for space-time prediction but in our case we do not due to the two assumptions we made when we separated spatial and temporal transitions (see Equation (2)). With those assumptions, the temporal transition density $\psi_n^{(2)}(t_{n+1}|T_n)$ is invariant for all locations within the study region at any given instant t_{n+1}. To present the predictions made by our model as a series of density maps over the study region indexed on time instants, only the relative magnitudes of the density estimates are relevant at any given instant. In fact the reader will see later that using relative magnitudes is essential to our approach to model evaluation and comparison. Therefore, we can safely ignore any components in the transition density model that do not depend on locations. These also include the normalizing constants in Equations (3) and (4), respectively.

---

[3] Technically, we have assumed that no two events happen at the same location or at the same time.

Intuitively, the number C of the clusters in the key feature space corresponds to the number of distinct sets of preferences. Unless we have this information *a priori* (e.g., crime analysts may tell us how many groups of offenders are likely to be represented by the data), we have to "discover" it from the data. Technically, the purpose of partitioning feature data is to effectively accommodate local covariance structures in the component density models that we will see momentarily. To accomplish this first task, we use a hierarchical clustering algorithm to generate partitions and employ a "stopping" rule to determine which partition is the "best." For a data set of n instances, a hierarchical clustering algorithm generates a succession of n partitions $P_0, P_1, ..., P_{n-1}$, where $P_0, P_1, ..., P_{n-1}$ contain n, n-1, ..., 1 clusters, respectively. It merges two "closest" clusters in P_j to generate P_{j+1} at each step. What we mean by "closest" obviously depends on the definition of cluster-to-cluster distance. This definition distinguishes different flavors of the algorithm. We will not delve into the details and the interested reader is referred to Everitt (1991) for a quick introduction. The "stopping" rule that we use is either the one proposed by Mojena (1977), which is essentially a "selection" rule (in the sense that it selects a partition from the complete sequence $P_0, P_1, ..., P_{n-1}$ after they are all generated rather than signaling a stop to the algorithm at an appropriate stage, say, P_j) or a revised version of it as stated below. Let α_j be the shortest distance between any two clusters in the partition P_j ($j = 0,1,..,n-1$). Then revised rule is to stop merging clusters further and select the first partition P_j satisfying

$$\alpha_{j+1} > \overline{\alpha}_j + k \cdot s_{\alpha_j} \qquad (5)$$

where $\bar{\alpha}_j$ and s_{α_j} are the mean and unbiased standard deviation of $\alpha_0, \alpha_1, ..., \alpha_j$, and the constant k is usually set to 1.25, as recommended by Milligan and Cooper (1985). When n is large, we find that this revised rule yields similar result to Mojena's original proposal. The rationale of these rules is to look for significant "jump" in the α series.

We consider two classes of models for estimating the first order spatial transition density. Both classes play an important role in modeling data that are believed to come from multiple underlying categories and sources. The first class is called *finite mixture distributions* (e.g., Everitt and Hand, 1981; Titterington et al., 1985; McLachlan and Basford, 1988). These distributions are superpositions of (usually simpler) component distributions. A finite mixture probability density function (or mass function in the case of discrete sample space) has the form

$$f(\mathbf{x}; \boldsymbol{\pi}, \Theta) = \sum_{j=1}^{C} \pi_j f_j(\mathbf{x}; \boldsymbol{\theta}_j) \qquad (6)$$

where $\pi_j > 0$, $j = 1,2,...,C$, $\pi_1 + \pi_2 + ... + \pi_C = 1$, $\boldsymbol{\pi} = [\pi_1 \quad ... \quad \pi_C]'$, $\Theta = [\boldsymbol{\theta}_1 \quad ... \quad \boldsymbol{\theta}_C]$. $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ is the jth *component density* with the set $\boldsymbol{\theta}_j$ of parameters and $\pi_1, \pi_2, ..., \pi_C$ are *mixing weights*. Θ is the collection of all *component parameters*. To fit a finite mixture distribution one needs to find the number C of component densities first. In our case this is done by task (1) — partitioning the feature data $\{\mathbf{x}_i : i = 1,2,..,n\}$.

Two aspects need to be addressed further in order for us to generate a density estimate by (6). First, further assumptions need to be made on the functional form of the component densities $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ $(j = 1,2,...,C)$. For a continuous feature space (where all features are continuous variables) we use Gaussian mixture models (GMM), where $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$, $j = 1,2,...,C$, are postulated as multivariate Gaussian. In the discrete case, we

fit the data with a class of Latent Class Models (LCM) (see Everitt, 1984), where we have assumed that the categorical feature variables are independent and the outcomes of each variable are also independent. For the situation where mixed variable types are present, it is trivial to combine GMM and LCM provided that the numeric dimensions are independent of the categorical ones. Second, we need an algorithm to estimate the set of parameters $\Theta = [\theta_1 \quad ... \quad \theta_C]$. We use a numeric maximum likelihood algorithm known as Expectation-Maximization (EM) algorithm (see, for example, Dempster, Laird and Rubin, 1977). Basically, the algorithm first calculates these parameters with respect to the clusters in the feature space partition, and then updates them iteratively until the log likelihood $L = \sum_{i=1}^{n} \log f(\mathbf{x}_i; \boldsymbol{\pi}, \Theta)$ converges to a stationary point.

The second class of techniques that we use to estimate the first order spatial transition density are nonparametric models and was introduced by Marchette et al. (1996). They are collectively called *filtered kernel estimators* (FKE) and take the form

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{\rho_j(\mathbf{x}_i)}{|\mathbf{H}_j|} K\left(\mathbf{H}_j^{-1}(\mathbf{x} - \mathbf{x}_i)\right) \tag{7}$$

where $K(\cdot)$ is termed a kernel function, $\mathbf{H}_j$, $j = 1,2,...,C$, are C $p \times p$ nonsingular *local bandwidth matrices* and $\rho_j(\mathbf{x})$, $j = 1,2,...,C$, which satisfy

$$0 \le \rho_j(\mathbf{x}) \le 1 \quad \text{and} \quad \sum_{j=1}^{C} \rho_j(\mathbf{x}) = 1 \tag{8}$$

for all $\mathbf{x}$, are *filtering functions*. Local bandwidth matrices contain posterior parameter settings that enforce localized smoothness for locally varied regions of the support of the true density. The filtering functions can be interpreted as prior weights over variations of local smoothness. We only consider a special case of (6) for our purpose where we set

$\mathbf{H}_j = \text{diag}[h_{j1} \quad ... \quad h_{jp}]$, $j = 1,2,...,C$, where h_{jl} $(j = 1,2,...,C, l = 1,2,..,p)$ is a local bandwidth for the lth dimension $[\mathbf{x}]_l$ of the jth locally varied region. We call these special class of estimators *filtered product kernel (FPK) estimators*. The underlying assumption for FPK estimators is that all dimensions are mutually independent.

In this paper we assume that the kernel function is standard multivariate Gaussian. To generate a density estimate by (7), we need to specify the filtering functions as well as the local bandwidths. Suppose the data $\{\mathbf{x}_i : i = 1,2,..,n\}$ have been partitioned into C clusters $\Omega_1, \Omega_2, ..., \Omega_C$. Let n_j be the number of instances in cluster Ω_j. We derive the filtering functions in one of the following two ways:

- Fit a finite mixture model $g(\mathbf{x}) = \sum_{j=1}^{C} \pi_j g_j(\mathbf{x})$ to the data. Set

$$\rho_j(\mathbf{x}) = \pi_j g_j(\mathbf{x}) / g(\mathbf{x}), \quad j = 1,2,...,C. \tag{9}$$

- Let the indicator $\mathbf{1}_{\{\mathbf{x} \in \Omega_j\}}$ be 1 if $\{\mathbf{x} \in \Omega_j\}$ and 0 otherwise. Set

$$\rho_j(\mathbf{x}) = \mathbf{1}_{\{\mathbf{x} \in \Omega_j\}}, \quad j = 1,2,...,C. \tag{10}$$

We term the FPK estimators with the filtering functions defined by (10) *weighted product kernel (WPK) estimators*. The local bandwidths are estimated by using local data in each cluster. To wit,

$$\hat{h}_{jl} = \left(\frac{4}{p+2}\right)^{1/(p+4)} \hat{\sigma}_{jl} n_j^{-1/(p+4)}, \; l = 1,2,...,p, \; j = 1,2,...,C. \tag{11}$$

where $\hat{\sigma}_{jl}$ is the standard deviation of the lth variable $[\mathbf{x}]_l$ estimated from $\{[\mathbf{x}_i]_l : \mathbf{x}_i \in \Omega_j, i = 1,2,...,n\}$. Notice that these bandwidth estimates are optimal in the AMISE sense assuming we were to fit Gaussian product kernel estimators to the local

data sets which are in fact samples of multivariate Gaussian distributions (see Scott, 1992).

When we use either finite mixture or filtered kernel estimators to model first order spatial transition density, models of distinct local structures are readily available. Spatial interaction probabilities are estimated from these "local" models. When a finite mixture distribution is involved, spatial interaction probabilities are given as

$$\Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\} = \pi_j f_j(\mathbf{x}_{n+1};\boldsymbol{\theta}_j) / f(\mathbf{x}_{n+1};\boldsymbol{\pi},\boldsymbol{\Theta}), \quad j = 1,2,...,C. \tag{12}$$

When a filtered kernel estimator is used, spatial interaction probabilities are given as

$$\Pr\{x_{n+1} \in \chi_n^{(j)}\} = \hat{f}_j(\mathbf{x}_{n+1}) / \hat{f}(\mathbf{x}_{n+1}), \quad j = 1,2,...,C \tag{13}$$

where

$$\hat{f}_j(\mathbf{x}_{n+1}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\rho_j(\mathbf{x}_i)}{|\mathbf{H}_j|} K\left(\mathbf{H}_j^{-1}(\mathbf{x}_{n+1} - \mathbf{x}_i)\right), \quad j = 1,2,...,C. \tag{14}$$

The third task on our list is to model second order spatial transition densities. The models we choose for these densities maintain certain continuity in parallel with the ordering of inter-event geographic distances and/or that of inter-event temporal distances. Such orderings reflect some additional assumptions on site selection behavior. First, given that two geographic locations have the same set of feature values, it is often reasonable to postulate that *event initiators are in favor of the geographically closer location to start the next event.* This assumption is supported by the "journey to crime" theory in criminology. In view of this assumption, a model of spatial interaction should give decreasing weight to past events with increasing distance to the location of interest. Another behavioral assumption that may hold true for certain scenarios (e.g., serial crimes of certain type) is that *event initiators tend not to wait long before they act again.*

A model incorporating this assumption should weigh the impacts of past events on future events according to their "ages". The more recently an event occurred, the higher weight it gets. Two models developed by Fiksel (1984), known as the order model and the instant model, both incorporate the "journey to event" assumption, while the instant model also take into account the assumption regarding "lingering period to resume act". We give these models below.

Let the number of data units in cluster j be m. Let $D_n^{(j)}=\{\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_m\}$ and $T_n^{(j)}=\{t_1,t_2,...,t_m\}$ where $t_1<t_2<...<t_m$ and $\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_m$ are ordered by $t_1,t_2,...,t_m$. Adapting Fiksel's order model to our case, we postulate the following function for the second-order spatial transition density for cluster j

$$\psi_n^{(12)}\left(\mathbf{s}\middle|D_n^{(j)},T_n^{(j)},t\right)=\varphi_m\left(\mathbf{s}\middle|\mathbf{s}_1,...,\mathbf{s}_m\right)=\frac{\lambda^2}{2\pi m}\sum_{i=1}^{m}e^{-\lambda\|\mathbf{s}-\mathbf{s}_i\|} \qquad (15)$$

where $t>t_m$ is a future event's time of occurrence and $\|\mathbf{s}-\mathbf{s}_i\|$ the distance from that future event's location $\mathbf{s}$ to an older event location $\mathbf{s}_i$ $(i=1,2,...,m)$. This is called an order model since only the temporal order of the events is considered. The instant model actually utilizes the values of the series $t_1,t_2,...,t_m$. Based on this model, we postulate that the second order spatial transition density for cluster j takes on the form

$$\psi_n^{(12)}\left(\mathbf{s}\middle|D_n^{(j)},T_n^{(j)},t\right)=\eta_m\left(\mathbf{s}\middle|\mathbf{s}_1,...,\mathbf{s}_m,t_1,...,t_m,t\right)=\frac{\lambda^2}{2\pi\sum_{i=1}^{m}e^{-\tau(t-t_i)}}\sum_{i=1}^{m}e^{-\lambda\|\mathbf{s}-\mathbf{s}_i\|-\tau(t-t_i)}. \qquad (16)$$

For both (15) and (16), we can numerically solve for the maximum likelihood estimates of the parameters (i.e., λ in (15), λ and τ in (16)). The interested reader is referred to Fiksel (1984).

The fourth and last task on our list is to estimate the geographic-space feature density when appropriate and possible. In general, this needs sampling over the study region. For example, we may obtain feature values for the locations on a regular grid over the study region. We may then fit a density function to these sample values using either finite mixture or filtered kernel method. This is the approach we take in the example that we give in Section 4.

3.3. Feature selection

So far we have assumed that our initial feature set coincides with the key feature set. By doing so, we have skipped the feature selection step to be described in this subsection. A feature selection problem can generally be specified by a triplet (F, c, s), where F is the *initial feature set*, c a *criterion function* defined for subsets of F, and s is a *subset search or selection procedure*. For the selection procedure, oftentimes we can just compare the scores of individual features and rank them accordingly. This is known as feature ranking and will be the approach we apply to the example in next section. If we select a subset of features based on scores of individual features, the underlying assumption is that these features are independent. Our emphasis in this paper is on feature selection criteria. In particular, we will see how we can exploit the point pattern in feature space to discover which feature or features are most predictive.

In Section 2, we have said that we should observe a distinct clustering (or *cohesive*) pattern consisting of small and well-separated cliques in the key feature space. The question then becomes how to gauge the cohesiveness of a point pattern in the feature subspace specified by a given set of features. The cohesiveness of this point pattern corresponds to the "goodness" of the feature set. The most straightforward

approach is to use some clustering algorithm to partition data units into the "best number" of clusters in the feature subspace and then examine inter-cluster separation and intra-cluster spread based on the cluster means and the cluster covariances. Friedman and Rubin (1967) gave several criteria in this class. An obvious problem with these criteria is that they require that we partition the data into the "best number" of clusters in the feature subspace defined by the feature subset to be evaluated. This partitioning problem is frequently not a trivial one to solve.

In this paper we look at another class of cohesiveness measures that do not require any partitioning in advance. These measures are functions of inter-event distances (or similarities). They all account for a basic characteristic of cohesive structures: It is nearly always the case that the distance between a pair of events is either very small or very large (i.e., the two events are either very similar or very dissimilar); rarely are two events separated by average inter-event distance. We define one of such measures in the following. Let d_{ij} be the distance between two data points i and j in the feature subspace defined by the feature subset to be evaluated. We transform the distance into the similarity s_{ij} by letting

$$s_{ij} = \frac{1}{1 + \alpha d_{ij}}, \tag{17}$$

where $\alpha = 1/\bar{d}$ and $\bar{d}$ is the averaged inter-event distance. Define the Gini index between these two events as follows.

$$g_{ij} = 4 s_{ij}(1 - s_{ij}). \tag{18}$$

Notice that g_{ij} attains its maximum of 1.0 when $s_{ij} = 0.5$ (or $d_{ij} = \bar{d}$) and its minimum of 0.0 when $s_{ij} \to 0.0$ (or $d_{ij} >> 1$) or $s_{ij} = 1.0$ (or $d_{ij} = 0$). For a data set of n events, the averaged Gini index defined by (19) is a suitable cohesiveness measure.

$$I_g = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} g_{ij}}{n(n-1)}. \qquad (19)$$

Smaller I_g corresponds to higher level of cohesiveness of the point pattern or a better set of features.

We note several caveats when using I_g for feature selection in practice. First, I_g is not intended for addressing a single-cluster pattern. It is only evaluated if every dimension of the data set for feature selection exhibits *enough* variation in values relative to the full range of that dimension over the entire region of interest. Operationally, we check each dimension of the data set and exclude the features that do not exhibit enough variation. Domain knowledge is critical to determine whether these features are among the most predictive ones or the most irrelevant ones to the problem at hand. Ideally, in the criminal event scenario, we could have very predictive features with correlations with event occurrence that are almost deterministic. If we found such a feature (judging from an analyst's experience, for instance), we could directly mark out the locations that have feature values within the feature's observed range as "hot spots." It is not necessary to include such a feature in a multivariate density estimation model because its contribution to the density score will dominate the contributions of other selected features anyway[4]. Second, any cohesive pattern (other than a single-cluster structure) as signaled by small I_g could imply strong correlation between the events showing the pattern and the features I_g

serves to evaluate, but it could also merely reflect the joint distribution of the features over the entire study region, or the prior distribution. In other words, the I_g score obtained for a set of features based on an event feature data set could be severely skewed by the prior distribution of these features. To single out the effect that the set of features has on the event of interest, the I_g score should be adjusted to eliminate the influence of the prior feature distribution as long as that distribution is not uniform.

Suppose that we can sample feature variables at locations on a regular grid, which is fine enough to represent all the locations within the study region. As opposed to the *event feature data set* we use to calculate an unadjusted I_g, we call the set of the feature values at the grid points the *prior feature data set*. We calculate an I_g score for the prior feature data set and let the score be I_g^p. Then we may adjust the I_g score for an event feature data set (or a feature subset to be evaluated) as follows.

$$\text{Adjusted } I_g = (\text{unadjusted}) \; I_g \; / \; I_g^p \tag{20}$$

4. Model Evaluation

In this section, we give a real-world application of our proposed transition density model. Based on this application, we compare statistically the results of our model with those obtained from the traditional space-time prediction methodology of using "hot-spots". We begin with a description of this traditional approach.

Recall that the transition density model we described in the last section simultaneously considers times, locations, and features of spatial-temporal events. Traditional space-time prediction models do not include feature data and criminal

[4] Technically, as long as the observed variance of a feature is not zero, the inclusion of the feature in a density estimation model will not cause singularity or infinite density score.

preferences over this feature data. The most sophisticated law enforcement agencies model criminal incidents as "hot-spots" or clusters in space and time. They then predict that future incidents will continue to occur in the observed or discovered clusters.

A variety of methods are used to perform this clustering. To be fair in this comparison, we will use exactly the same techniques for clustering in space and time as we use in our model. Hence, the only difference will be that our model also includes clustering and preference discovery in feature space. As a result, our evaluation with the comparison models (models without feature data) will tell us if we can gain any predictive power from modeling criminal preferences in feature space as well as in geographic space.

A formal description of the comparison models is as follows. Ignoring the feature data, a comparison model predicts the likelihood of the occurrence of a future event $(\mathbf{s}_{n+1}, t_{n+1})$ based on the locations and times of past events $(\mathbf{s}_1, t_1)$, $(\mathbf{s}_2, t_2)$, ... $(\mathbf{s}_n, t_n)$, $t_0 = 0 < t_1 < t_2 < \ldots < t_n < t_{n+1}$. The quantity of interest is the density function $\psi_n(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n)$, where $T_n = \{t_1, t_2, \ldots, t_n\}$ and $D_n = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. Under the assumption that the occurrence of events over time and space are independent, the comparison model is specified by

$$\psi_n(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n) = \psi_n^{(1)}(\mathbf{s}_{n+1} | D_n) \cdot \psi_n^{(2)}(t_{n+1} | T_n). \qquad (21)$$

In parallel with our model, we term $\psi_n^{(1)}(\mathbf{s}_{n+1} | D_n)$ the *spatial transition density* and $\psi_n^{(2)}(t_{n+1} | T_n)$ *temporal transition density*. However, unlike $\psi_n^{(1)}(\mathbf{s}_{n+1} | D_n, \chi_n, T_n, t_{n+1})$ in our model, $\psi_n^{(1)}(\mathbf{s}_{n+1} | D_n)$ only captures the event continuity in geographic space. In other words, $\psi_n^{(1)}(\mathbf{s}_{n+1} | D_n)$ assigns high densities only to the locations in the vicinity of old

event locations. Due to the reasons we stated for our model in the last section, the temporal transition density $\psi_n^{(2)}(t_{n+1}|T_n)$ can be safely ignored from the comparison model in our case.

The space-time events of interest in our application are both commercial and residential "breaking & entering" (B&E) incidents that occurred in Richmond, Virginia. A total of 579 such incidents happened between July 1, 1997 and August 31, 1997 and that is the time range for our study. These incidents are singled out primarily because they constitute a data set of reasonable size for rolling weekly and biweekly analyses. Table 1 summarizes the weekly counts of the B&E incidents in the study horizon. Notice that the crime rate rose to a steady level starting the second week of July and did not drop until the second to last week of August. Since the reason for the changes in crime rate is not clear, we choose not to use the data from the first week of July and the last two weeks of August for model building in the sequel.

The data associated with these incidents come from multiple sources. The locations and the times of the criminal incidents were originally stored in a central dBase database called Prism, and are made available to us through our project work with the Richmond Police. Figure 2 shows the locations of the B&E incidents on the map of Richmond. The subregions on the map are block groups, which are the smallest areas for which census counts are tallied. We consider three types of features related to B&E incidents. The demographic and consumer expenditure features data are converted from the 1997 estimates of certain census categories recorded in "CensusCD+maps" (1998). The distances from crime locations to geographic landmarks are generated by the GIS component of the ReCAP system, a crime-fighting decision support software being built

by the researchers at the University of Virginia. The three types of feature variables are listed in figures 3, 4, and 5, respectively. We assume that the feature values at any given location in the study region remain unchanged within the study horizon. Simply put, this means we have a static study region. Given the nature of our initial features and the fact that our study horizon spans only two months (i.e., July and August of 1997), this seems to be a rather safe assumption.

Week	No. of Incidents
July 1 – 6	50
July 7 – 13	74
July 14 – 20	71
July 21 – 27	72
July 28 - August 3	68
August 4 – 10	69
August 11 – 17	72
August 18 – 24	54
August 25 – 31	49

Table 1. Weekly counts of Breaking and Entering criminal incidents between July 1, 1997 and August 31, 1997 in Richmond, Virginia.

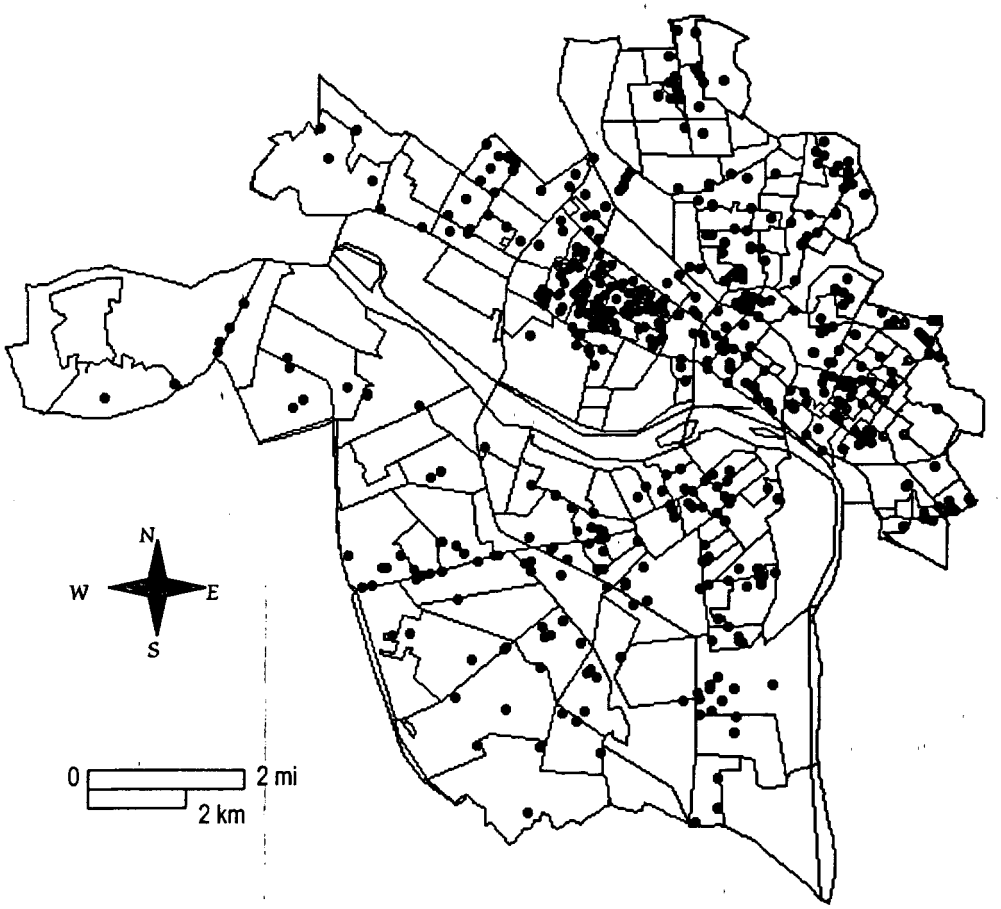

Figure 2. Breaking and Entering criminal incidents between July 1, 1997 and August 31, 1997 in Richmond, Virginia.

Feature	Description
Population, General	
POP_DST	Per square mile (psm) population
HH_DST	Households psm
FAM_DST	Families psm
MALE_DST	Male population psm
FEM_DST	Female population psm
Work Force	
CLS12_DST	Private wage and salary workers psm
CLS345_DST	Government workers psm
CLS67_DST	Self-employed and unpaid family workers psm
Income	
PCINC_97	Per capita annual income
MHINC_97	Median annual household income
AHINC_97	Average annual household income
Householder Age	
AGEH12_DST	Households with householder under 34 years of age psm
AGEH34_DST	Households with householder between 35 to 54 years of age psm
AGEH56_DST	Households with householder above 55 years of age psm
Household Size	
PPH1_DST	1 person households psm
PPH2_DST	2 person households psm
PPH3_DST	3-5 person households psm
PPH6_DST	6 or more person households psm
Housing Structure	
HSTR1_DST	Occupied structures with 1 unit detached psm
HSTR2_DST	Occupied structures with 1 unit attached psm
HSTR3_DST	Occupied structures with 2 units psm
HSTR4_DST	Occupied structures with 3-9 units psm
HSTR6_DST	Occupied structures with 10+ units psm
HSTR9_DST	Occupied trailers psm
HSTR10_DST	Other occupied structures psm
Housing, Miscellaneous	
HUNT_DST	Housing units psm
HUNT_PC	Per capita housing units
OCCHU_DST	Occupied housing units psm
OCCHU_PC	Per capita occupied housing units
VACHU_DST	Vacant housing units psm
MORT1_DST	Owner occupied housing units with mortgage psm
MORT2_DST	Owner occupied housing units without mortgage psm
COND1_DST	Owner occupied condominiums psm
OWN_DST	Owner occupied units psm
RENT_DST	Renter occupied units psm

Table 2. Demographic features.

Feature	Description
APPAREL_PH	Per household annual expenditure (phae) on apparel and footwear
APPAREL_PC	Per capita annual expenditure (pcae) on apparel and footwear
ALC_TOB_PH	Phae on alcohol beverages, tobacco and smoking
ALC_TOB_PC	Pcae on alcohol beverages, tobacco and smoking
EDU_PH	Phae on education
EDU_PC	Pcae on education
ET_PH	Phae on entertainment
ET_PC	Pcae on entertainment
FOOD_PH	Phae on food
FOOD_PC	Pcae on food
MED_PH	Phae on drugs, health insurance, medical services and supplies
MED_PC	Pcae on drugs, health insurance, medical services and supplies
HOUSING_PH	Phae on household furnishings, operations, and shelter
HOUSING_PC	Pcae on household furnishings, operations, and shelter
P_CARE_PH	Phae on personal care, personal insurance and pension
P_CARE_PC	Pcae on personal care, personal insurance and pension
REA_PH	Phae on reading
REA_PC	Pcae on reading
TRANS_PH	Phae on public transportation, vehicle purchase and maintenance
TRANS_PC	Pcae on public transportation, vehicle purchase and maintenance

Table 3. Consumer expenditure features.

Feature	Description
D_SCHOOL	Distance to the nearest school
D_HIGHWAY	Shortest distance to the nearest highway
D_HOSPITAL	Distance to the nearest hospital
D_CHURCH	Distance to the nearest church
D_PARK	Distance to the nearest park

Table 4. Distance features.

To select the key feature set, we calculate the I_g score for each initial feature (shown in tables 2, 3 and 4) with the feature data pertaining to the B&E incidents between July 7, 1997 and July 20, 1997 (i.e., the event feature data set for feature selection). We then adjust the score with the I_g score obtained based on the feature data pertaining to 2517 locations placed evenly over the Richmond map (i.e., the prior feature data set for feature selection). This regular grid of 2517 sample points is also involved in constructing model comparison statistics that we will describe momentarily. The results are reported in tables

5, 6 and 7. It is noted here that before we computed the I_g scores, we have first examined the ratio of the observed range (calculated from the event feature data set) to the full range (calculated from the prior feature data set) for each initial feature to see whether there are any features that do not exhibit enough variations in the event feature data set. It turns out that this ratio is greater than 0.2 for every initial feature in our example. We deem this an indicator that there is enough variation in every feature dimension.

Feature	IL	Adj IL		Feature	IL	Adj IL
Population, General				*Housing Structure*		
FAM_DST	***0.795109***	***0.971294***		HSTR9_DST	0.209613	0.430049
FEM_DST	0.780887	1.017172		HSTR6_DST	0.578788	0.971377
HH_DST	0.766205	1.019083		HSTR1_DST	0.779776	1.037161
POP_DST	0.77807	1.022192		HSTR4_DST	0.603965	1.095686
MALE_DST	0.77391	1.037627		HSTR10_DST	0.511243	1.171066
Work Force				HSTR2_DST	0.513737	1.33525
CLS12_DST	0.762812	0.99573		HSTR3_DST	0.442481	1.543366
CLS67_DST	0.71836	1.013683		*Housing, Miscellaneous*		
CLS345_DST	0.755043	1.020015		COND1_DST	0.28449	0.249759
Income				OCCHU_DST	0.766194	1.019019
PCINC_97	0.746605	1.093547		MORT1_DST	0.778619	1.034395
MHINC_97	0.74147	1.100745		HUNT_DST	0.764804	1.035979
AHINC_97	0.700613	1.16912		OWN_DST	0.77991	1.051672
Householder Age				RENT_DST	0.691134	1.054123
AGEH12_DST	0.689906	0.979065		OCCHU_PC	0.755908	1.070385
AGEH56_DST	0.758949	1.017699		HUNT_PC	0.762469	1.072405
AGEH34_DST	0.776586	1.047537		MORT2_DST	0.74747	1.075255
Household Size				VACHU_DST	0.689763	1.088101
PPH1_DST	0.698101	0.999252				
PPH2_DST	0.774179	1.019169				
PPH3_DST	0.770058	1.019687				
PPH6_DST	0.648417	1.096216				

Table 5. Demographic features evaluation result.

Feature	I_g	Adj I_g	Feature	I_g	Adj I_g
Per Household			*Per Capita*		
P_CARE_PH	***0.778652***	***0.886927***	P_CARE_PC	0.804807	0.958234
TRANS_PH	0.748267	0.961544	EDU_PC	0.802809	0.978819
MED_PH	0.791697	0.969762	HOUSING_PC	0.806986	0.980284
ET_PH	0.789273	0.97886	APPAREL_PC	0.813909	0.99788
HOUSING_PH	0.697043	1.005566	ET_PC	0.816095	0.998878
REA_PH	0.784346	1.015941	TRANS_PC	0.821257	1.001076
APPAREL_PH	0.784296	1.018549	ALC_TOB_PC	0.816618	1.007928
EDU_PH	0.759107	1.02109	MED_PC	0.813172	1.012766
ALC_TOB_PH	0.784793	1.025226	FOOD_PC	0.804328	1.013596
FOOD_PH	0.748634	1.044432	REA_PC	0.798631	1.015429

Table 6. Consumer expenditure features evaluation result.

Feature	I_g	Adj I_g
D_HIGHWAY	***0.80264***	***0.99483***
D_PARK	0.798587	1.003996
D_SCHOOL	0.756689	1.0291
D_CHURCH	0.795715	1.032549
D_HOSPITAL	0.79801	1.036391

Table 7. Distance features evaluation result.

We select one feature from each table to form the key feature set. We do not select features from the same table so as to avoid strong correlation between any two features in the key feature set. Independence between features is an assumption for the versions of the transition density model to be calibrated and evaluated. The features that we pick based on adjusted I_g are FAM_DST (Families per square mile), P_CARE_PH (Per household annual expenditure on personal care, personal insurance and pension) and D_HIGHWAY (Shortest distance to the nearest highway). We bypass two features COND1_DST and HSTR9_DST which have lower adjusted I_g than FAM_DST for both technical and practical reasons. Technically, these two features have unusually low I_g scores on the prior feature data set (as compared with other features), which indicate that

the prior feature data set for either feature is highly clustered or the prior distribution of either feature is far from uniform. This intuitively makes sense since out of the 207 block groups in Richmond there are only several that have occupied trailer homes or owner occupied condominiums. Even with adjustment we still cannot completely eliminate the influence of the prior patterns on the event feature data for both features. This is reflected in their very low adjusted I_g scores. Practically, we eliminate these features because when working with crime analysts we find them unwilling to claim that the lack of trailer homes or condominiums is linked to higher rate of B&E incidents. However, further analysis is clearly recommended on this issue and our model easily supports this additional analysis.

Geo-mapping shows that the distribution of each selected feature roughly correlates to criminal event intensity as described in the following. Firstly, the intensity of B&E incidents is roughly proportional to family density. Second, how much on average a household within a region spends on personal care products and services is a reasonable indicator of disposable income with the block group. Most of the criminal incidents concentrate in the low to middle values of this attribute but not as much in the highest or lowest values. Lastly, areas close to highways are prone to B&E incidents. The fact that we have combined both residential and commercial B&E incidents may account for this. Other explanations relate to the opportunity to commit crimes provided by highways.

We evaluate three versions of our model against their counterparts comparison models. The three versions are named GMM, WPK and FPK. The GMM version of the proposed model uses Gaussian mixture models for estimating both the first order spatial

transition density and the geographic-space feature density. The GMM version of the comparison model also uses a Gaussian mixture model for estimating the spatial transition density. The WPK version replaces Gaussian mixture estimation with weighted product kernel estimation and the FPK version uses filtered product kernel estimation. We build the three versions of the proposed model on four training data sets and for each version we test it and compare it with the corresponding comparison model under three test scenarios – substituting training data back into the model (resubstitution), predicting out one week into the future (weekly prediction), and predicting out two weeks into the future (biweekly prediction). The training sets are the data sets associated with the B&E incidents that occurred during these four fortnights, July 7 to 20, July 14 to 27, July 21 to August 3, and July 28 to August 10, respectively. For every version of the proposed model, we always use the same set of key features that we just selected (based on the feature data of the incidents between July 7 and July 20). By doing so we assume that the preferences for initiating B&E incidents remain static during the entire period of July and August. Alternatively, one can repeat the feature selection process described in the last section on each training data set if one believes that the preferences are dynamic over this period.

To compare the performances of different models, we convert the density estimates into *percentile scores* which are on a common scale of 0 to 100. Suppose that we have placed over the study region a regular grid consisting of N points. In our case, $N = 2157$. Let $\mathbf{s}_i^g$ be the location of the ith grid point. Denote the density estimate (generated by either the proposed model or the comparison model) at an arbitrary location $\mathbf{s}$ as $d_{\mathbf{s}}$. The percentile score $p_{\mathbf{s}}$ at location $\mathbf{s}$ is defined by

$$p_s = (100/N)\sum_{i=1}^{N} \mathbf{1}\{d_s \geq d_{s_i^g}\} \quad (22)$$

where $\mathbf{1}\{d_s \geq d_{s_i^g}\}$ is 1 if $d_s \geq d_{s_i^g}$ and 0 otherwise. Assuming that the grid is fine enough to represent the study region well, percentile scores are nothing but re-scaled density estimates. The higher a percentile score at a specified location then the more likely it is that a new event will happen at that location.

Basic model evaluation statistics are given in terms of mean predicted percentile score and its standard deviation for three versions of the proposed model and three versions of the comparison model calibrated on the four aforementioned training data sets in tables 8, 10, 12, 14, respectively. The "best model" is referred to as the version of a model with the highest mean percentile score out of the three versions of that model. It is clearly seen from these tables that the proposed model outperforms the comparison model in every test scenario in terms of mean percentile score. But is this result statistically significant?

Two hypothesis tests are performed to answer this question. Assume that the test data set contains m incidents that occurred at the locations $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m$, respectively. For the incident at $\mathbf{s}_i$, let the percentile score given by the proposed model be $p_{\mathbf{s}_i}^p$ and that given by the comparison model be $p_{\mathbf{s}_i}^c$. Let δ be the probability that the proposed model outperforms the comparison model on a single prediction. We perform the hypothesis test

$$H_0: \delta = 0.5,$$
$$H_a: \delta > 0.5,$$

if the test statistic $\hat{\delta} > 0.5$; otherwise, we test the same null hypothesis against

$$H_a: \delta < 0.5.$$

The test statistic $\hat{\delta}$ for the first hypothesis test is given as follows.

$$\hat{\delta} = (1/m)\sum_{i=1}^{m} \mathbf{1}\{p_{s_i}^p > p_{s_i}^c\}. \quad (23)$$

The second hypothesis test is built around μ which denotes the mean of the difference between the percentile score given by the proposed model and that given by the comparison model on a single prediction. We perform the hypothesis test

$$H_0: \mu = 0,$$
$$H_a: \mu > 0,$$

if the test statistic $\hat{\mu} > 0$; otherwise, we test the same null hypothesis against

$$H_a: \mu < 0.$$

The test statistic $\hat{\mu}$ based on a test set of m incidents is straightforward. To wit,

$$\hat{\mu} = (1/m)\sum_{i=1}^{m} (p_{s_i}^p - p_{s_i}^c). \quad (24)$$

The standard deviation of the difference $q_{s_i} = p_{s_i}^p - p_{s_i}^c$ is estimated by

$$\hat{\sigma} = (1/(m-1))\sum_{i=1}^{m} (q_{s_i} - \hat{\mu})^2. \quad (25)$$

The results of these tests are reported in tables 9, 11, 13, and 15, in which "Prob.", "Mean" and "Std. Dev." correspond to $\hat{\delta}$, $\hat{\mu}$ and $\hat{\sigma}$, respectively. These tables show that

- for all but one comparison, our model statistically performs better than the comparison model at the 90% confidence level according to the result of at least one hypothesis test;
- for the one comparison that both hypothesis tests fail at the 90% confidence level ("Best vs. Best" under weekly prediction in Table 9), the performances of the two models are statistically indistinguishable since the two hypothesis tests are set up

against opposite alternative hypotheses but neither test can reject the null in favor of the alternative; and

- for well over half of the hypothesis tests, the null hypothesis is rejected with a smaller p-value under biweekly prediction than it is under weekly prediction, which indicates that the proposed model is able to capture the patterns of event occurrences over a longer term due to the addition of the feature space analysis.

Training set: July 7-20 (145 incidents)				
Resubstitution - Test set: July 7-20 (145 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	86.0255	15.0389	58.3042	21.0441
WPK	89.5119	12.3379	83.0167	16.9308
FPK	89.5346	12.2509	83.0167	16.9308
Best Model	FPK		WPK or FPK	
Weekly prediction - Test set: July 21-27 (72 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	76.2956	26.2846	56.4876	22.7824
WPK	75.9381	25.2531	73.9604	26.5926
FPK	75.8023	25.2659	73.9604	26.5926
Best Model	GMM		WPK or FPK	
Biweekly prediction - Test set: July 21-August 3 (140 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	75.8502	24.0831	56.9950	23.1292
WPK	74.3845	25.1521	72.5277	26.0815
FPK	74.1628	25.1669	72.5277	26.0815
Best Model	GMM		WPK or FPK	

Table 8. Basic statistics for models calibrated on July 7-20 data.

Training set: July 7-20 (145 incidents)							
Resubstitution - Test set: July 7-20 (145 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8828	9.2180	<0.0002	27.7213	26.2747	12.7046	<0.0002
WPK vs. WPK	0.9379	10.5468	<0.0002	6.4951	7.9222	9.8724	<0.0002
FPK vs. FPK	0.9103	9.8824	<0.0002	6.5179	8.0996	9.6901	<0.0002
Best vs. Best	0.9103	9.8824	<0.0002	6.5179	8.0996	9.6901	<0.0002
Weekly prediction - Test set: July 21-27 (72 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.7500	4.2426	<0.0002	19.8081	32.5387	5.1655	<0.0002
WPK vs. WPK	0.5833	1.4142	0.0793	1.9777	10.9967	1.5260	0.063
FPK vs. FPK	0.5972	1.6499	0.0495	1.8419	10.9029	1.4335	0.0764
Best vs. Best	0.4444	0.9428	0.1736	2.3352	19.3500	1.0240	0.1539
Biweekly prediction - Test set: July 21-August 3 (140 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.7286	5.4090	<0.0002	18.8552	31.1996	7.1507	<0.0002
WPK vs. WPK	0.5857	2.0284	0.0212	1.8568	7.8414	2.8018	0.0026
FPK vs. FPK	0.5857	2.0284	0.0212	1.6352	7.9976	2.4192	0.0078
Best vs. Best	0.4786	0.5071	0.305	3.3225	15.8084	2.4868	0.0064

Table 9. Hypothesis tests results for models calibrated on July 7-20 data.

Training set: July 14-27 (143 incidents)				
Resubstitution - Test set: July 14-27 (143 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	81.0975	21.1945	61.6399	24.9543
WPK	85.7064	15.7511	79.8256	19.9957
FPK	85.7670	15.5135	79.8256	19.9957
Best Model	FPK		WPK or FPK	
Weekly prediction - Test set: July 28-August 3 (68 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	76.3117	21.6247	59.2512	27.6379
WPK	72.6162	25.2771	70.1436	27.1039
FPK	72.2990	25.2911	70.1436	27.1039
Best Model	GMM		WPK or FPK	
Biweekly prediction - Test set: July 28-August 10 (137 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	73.5904	24.2488	57.5221	26.5259
WPK	72.0119	26.5243	69.7636	27.5075
FPK	71.8226	26.4933	69.7636	27.5075
Best Model	GMM		WPK or FPK	

Table 10. Basic statistics for models calibrated on July 14-27 data.

Training set: July 14-27 (143 incidents)							
Resubstitution - Test set: July 14-27 (143 incidents).							
	Test 1			Test 2			
Comparison	Prob. > z	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.7832	6.7736	<0.0002	19.4576	28.0027	8.3092	<0.0002
WPK vs. WPK	0.9021	9.6168	<0.0002	5.8808	7.5411	9.3255	<0.0002
FPK vs. FPK	0.9021	9.6168	<0.0002	5.9414	7.6988	9.2286	<0.0002
Best vs. Best	0.9021	9.6168	<0.0002	5.9414	7.6988	9.2286	<0.0002
Weekly prediction - Test set: July 28-August 3 (68 incidents).							
	Test 1			Test 2			
Comparison	Prob. > z	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8088	5.0932	<0.0002	17.0605	27.7049	5.0780	<0.0002
WPK vs. WPK	0.6029	1.6977	0.0446	2.4726	8.3534	2.4409	0.0073
FPK vs. FPK	0.5882	1.4552	0.0721	2.1553	8.5092	2.0887	0.0183
Best vs. Best	0.5441	0.7276	0.2327	6.1681	14.7758	3.4423	0.0003
Biweekly prediction - Test set: July 28-August 10 (137 incidents).							
	Test 1			Test 2			
Comparison	Prob. > z	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.7664	6.2368	<0.0002	16.0683	29.5662	6.3611	<0.0002
WPK vs. WPK	0.6204	2.8194	0.0024	2.2484	8.7033	3.0237	0.0013
FPK vs. FPK	0.5766	1.7942	0.0367	2.0590	8.7983	2.7392	0.0031
Best vs. Best	0.5182	0.4272	0.3336	3.8268	16.7383	2.6760	0.0037

Table 11. Hypothesis tests results for models calibrated on July 14-27 data.

Training set: July 21-August 3 (140 incidents)				
Resubstitution - Test set: July 21-August 3 (140 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	79.7758	19.6797	60.1351	26.3057
WPK	80.7356	19.0859	77.1133	21.0586
FPK	80.6822	18.9690	77.1133	21.0586
Best Model	WPK		WPK or FPK	
Weekly prediction - Test set: August 4-10 (69 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	73.3315	23.8760	54.3498	25.3288
WPK	69.3522	28.3111	67.2620	29.6937
FPK	69.2837	28.2384	67.2620	29.6937
Best Model	GMM		WPK or FPK	
Biweekly prediction - Test set: August 4-17 (141 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	77.1184	22.5256	55.7052	25.7259
WPK	72.7329	27.0081	71.6619	27.6472
FPK	72.5565	27.0017	71.6619	27.6472
Best Model	GMM		WPK or FPK	

Table 12. Basic statistics for models calibrated on July 21-August 3 data.

Training set: July 21-August 3 (140 incidents)							
Resubstitution - Test set: July 21-August 3 (140 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8286	7.7754	<0.0002	19.6407	27.0597	8.5881	<0.0002
WPK vs. WPK	0.8571	8.4515	<0.0002	3.6222	5.5286	7.7522	<0.0002
FPK vs. FPK	0.8571	8.4515	<0.0002	3.5689	5.8308	7.2421	<0.0002
Best vs. Best	0.8571	8.4515	<0.0002	3.6222	5.5286	7.7522	<0.0002
Weekly prediction - Test set: August 4-10 (69 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.7971	4.9358	<0.0002	18.9816	29.8703	5.2786	<0.0002
WPK vs. WPK	0.5652	1.0835	0.1401	2.0901	10.8363	1.6022	0.0548
FPK vs. FPK	0.5797	1.3242	0.0934	2.0216	10.9703	1.5308	0.063
Best vs. Best	0.5797	1.3242	0.0934	6.0695	19.2327	2.6214	0.0044
Biweekly prediction - Test set: August 4-17 (141 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8298	7.8320	<0.0002	21.4133	28.0072	9.0787	<0.0002
WPK vs. WPK	0.5319	0.7579	0.2236	1.0710	4.8907	2.6003	0.0047
FPK vs. FPK	0.5532	1.2632	0.1038	0.8946	4.9496	2.1463	0.0158
Best vs. Best	0.5603	1.4317	0.0764	5.4565	16.9000	3.8339	<0.0002

Table 13. Hypothesis tests results for models calibrated on July 21-August 3 data.

Training set: July 28-August 10 (137 incidents)				
Resubstitution - Test set: July 28-August 10 (137 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	78.9968	20.4379	44.4110	26.5077
WPK	80.4291	19.2876	75.5647	22.8157
FPK	80.4457	19.0202	75.5647	22.8157
Best Model	FPK		WPK or FPK	
Weekly prediction - Test set: August 11-17 (72 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	81.6696	20.4393	38.5341	25.9068
WPK	76.2355	25.0248	75.4734	24.9736
FPK	75.9855	25.0196	75.4734	24.9736
Best Model	GMM		WPK or FPK	
Biweekly prediction - Test set: August 11-24 (126 incidents).				
	Proposed Model		*Comparison Model*	
Model Type	Mean	Std. Dev.	Mean	Std. Dev.
GMM	80.9086	20.9195	40.4462	25.3919
WPK	76.7429	23.8933	75.3779	24.0492
FPK	76.5105	23.9671	75.3779	24.0492
Best Model	GMM		WPK or FPK	

Table 14. Basic statistics for models calibrated on July 28-August 10 data.

Training set: July 28-August 10 (137 incidents)							
Resubstitution - Test set: July 28-August 10 (137 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8394	7.9455	<0.0002	34.5858	38.5808	10.4927	<0.0002
WPK vs. WPK	0.8905	9.1416	<0.0002	4.8644	8.3799	6.7944	<0.0002
FPK vs. FPK	0.8321	7.7747	<0.0002	4.8810	8.6396	6.6126	<0.0002
Best vs. Best	0.8321	7.7747	<0.0002	4.8810	8.6396	6.6126	<0.0002
Weekly prediction - Test set: August 11-17 (72 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8889	6.5997	<0.0002	43.1356	36.0015	10.1667	<0.0002
WPK vs. WPK	0.5972	1.6499	0.0495	0.7620	5.9915	1.0792	0.1401
FPK vs. FPK	0.6111	1.8856	0.0294	0.5121	5.9684	0.7280	0.2327
Best vs. Best	0.5278	0.4714	0.3192	6.1962	17.9847	2.9234	0.0018
Biweekly prediction - Test set: August 11-24 (126 incidents).							
	Test 1			Test 2			
Comparison	Prob.	z-Statistic	p-Value	Mean	Std. Dev.	z-Statistic	p-Value
GMM vs. GMM	0.8968	8.9087	<0.0002	40.4623	37.0444	12.2607	<0.0002
WPK vs. WPK	0.6111	2.4944	0.0064	1.3650	9.8093	1.5620	0.0594
FPK vs. FPK	0.6111	2.4944	0.0064	1.1326	9.8257	1.2939	0.0985
Best vs. Best	0.5238	0.5345	0.2981	5.5306	18.7787	3.3059	0.0005

Table 15. Hypothesis tests results for models calibrated on July 28-August 10 data.

Density maps generated by the three versions of the proposed model built on the training data of the 145 incidents between July 7 and July 20 are given in Figures 3. The criminal incidents occurring within the immediate following week or two weeks (i.e., the test sets) are plotted on the density maps to enable visual examination of how well the proposed model performs under weekly or biweekly prediction scenario. Similar density maps can be generated for the models built on other training data sets. It is easily seen on these maps that most of the test incidents indeed happened around the predicted "hot spots" (i.e., predicted high-density areas). Also by visual inspection, the GMM version of the proposed model seems to have captured more details than the WPK version and the FPK version in all four figures. This is confirmed in Tables 8, 10, 12 and 14 where the GMM version is indeed picked as the "best model" for every weekly or biweekly prediction scenario. The WPK and FPK versions seem to have equivalent performances. The density maps obtained for these versions look smoother than those obtained for the GMM version.

Figure 3. GMM (upper), WPK (middle) and FPK (lower) versions of the proposed model calibrated on July 7-20 data and tested on July 21-27 data (left) and July 21-August 3 data (right).

5. Conclusion

The development of predictive models of criminal activity is of tremendous value to law enforcement. The use of these models in support of tactical decision making in law enforcement is obvious: the better we forecast criminal activity then the better we can allocate law enforcement resources to combat it. However, the usefulness and significance of these models goes beyond tactical decision making. They effectively support community policing, problem-oriented policing, and cooperation among agencies.

In this paper, we have described a newly developed space-time prediction model and evaluated it on real-world data sets from the domain of regional crime analysis. The presented model is shown to be more effective than the traditional "hot-spot" methods, especially for predicting the occurrence of space-time events characteristic of human intelligence and preferences, as exemplified by the Richmond breaking and entering incidents. Distinctive from other methods in the literature, our modeling approach

- accommodates all measurable features useful for prediction,
- identifies which of the features have the most predictive or explanatory power, and
- generates probability density estimates over space and time for the occurrence of future events.

Specific to the law enforcement domain, this approach provides the basis for theory development, since it shows how community and law enforcement data relate over space and time. It also provides a vehicle for theory evaluation or testing, since it can show which theoretical relationships lead to accurate predictions and which do not. For

instance, for the Richmond Breaking and Entering crime application, we have found that such features as family density, disposable income (as indicated by per household personal care expenditure), and proximity to highways could jointly play a role in crime initiation decisions. The proposed model quantifies the form of correlation between these features and occurrence of B&E incidents.

Obviously, the applicability of our approach to preference discovery is not confined to law enforcement. For example, in military actions, one may want to predict the future location of an enemy target (e.g., a tank) moving over terrain based on its past locations (observed over predefined sampling intervals) and terrain features. In an urban development, developers are interested in predicting consumer behavior toward a new shopping mall using data from past behavior toward existing malls. They would also use data regarding surrounding neighborhoods and the physical infrastructure in the area (e.g., major highways, schools, and bridges). In this sense, our model provides a generic framework for space-time event forecasting.

References

Amir, M. (1971). *Patterns in Forcible Rape*. University of Chicago Press, Chicago.

Baldwin, J. and Bottoms, A. (1976). *The Urban Criminal: A Study in Sheffield*. Tavistock Publications, London.

Brantingham, P. and Brantingham, P. (1975). Spatial patterns of burglary. *Howard Journal of Penology and Crime Prevention*, **14**, 11-24.

Brantingham, P. and Brantingham, P. (1984). *Patterns in Crime*. Macmillan Publishing Company, New York.

Capone, D. and Nichols, W. (1976). Urban structure and criminal mobility, *American Behavioral Scientist*, **20**, 199-213.

CensusCD+maps, Version 2.0 (1998). GeoLytics, East Brunswick, NJ.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society, B*, **39**, 1-38.

Everitt, B. S. (1984). *An Introduction to Latent Variable Models*. Chapman and Hall, London.

Everitt, B. S. (1991). *Cluster Analysis*, 3rd Ed. Edward Arnold, London.

Everitt, B. S. and Hand, D. J. (1981). *Finite Mixture Distributions*. Chapman and Hall, London.

Fiksel, T. (1984). Simple spatial-temporal models for sequences of geological events. *Elektronische Informationsverarbeitung und Kybernetik*, **20**, 480-487.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of American Statistical Association*, 1159-1178.

LeBeau, J. L. (1987). The journey to rape: Geographic distance and the rapist's methods of approaching the victim. *Journal of Police Science and Administration*, **15**, 129-136.

Marchette, D. J., Priebe, C. E., Rogers, G. W. and Solka, J. L. (1996). Filtered kernel density estimation. *Computational Statistics*, **11**, 95-112.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, Inc., New York.

Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159-179.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal*, **20**, 359-363.

Molumby, T. (1976). Patterns of crime in a university housing project. *American Behavioral Scientist*, **20**, 247-259.

Newman, O. (1972). *Defensible Space: Crime Prevention Through Urban Design*. Macmillan, New York.

Repetto, T. A. (1974). *Residential Crime*. Ballinger, Cambridge, MA.

Rossmo, D. K. (1993). Target patterns of serial murders: A methodological model. *American Journal of Criminal Justice*, **17(2)**, 1-21.

Rossmo, D. K. (1994). Targeting victims: Serial killers and the urban environment. In *Serial and Mass Murder: Theory, Research, and Policy*, ed. by T. O'Reilly-Flemming and S. Egger, University of Toronto Press, Toronto.

Scarr, H. A. (1973). *Patterns in Burglary*, 2nd Ed., U.S. Department of Justice, Washington, D.C.

Scott, D. W. (1992). *Multivariate Density Estimation.* Wiley, New York.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Spatial-Temporal Point Process Models for Criminal Event Prediction

Donald E. Brown, University of Virginia
Steven H. Kerchner, University of Virginia

Abstract

An important need in crime prevention is to predict the likelihood that criminal incidents occur at specified locations within a geographic region and a specified time range, based on historical incident records. A model for predicting the probability of occurrence of spatial-temporal random events was developed based on the theory of point patterns. This model views the available data as a realization of a marked space-time shock point process, and the prediction problem as the estimation of the space-time transition density of the process. In contrast to traditional space-time prediction models, this model incorporates richly informative observed event characteristics or features into space-time prediction. Our previous model assumes temporally homogeneous data and thus excludes all temporal features. We describe our extension to the model that incorporates temporal feature heterogeneity by changing the temporal transition density calculation. Test results comparing this model with traditional methods for predicting hot spots show that the model with temporal features outperforms other approaches in some cases, but that use of temporal features is only effective in data sets that display temporal patterns.

Table of Contents

Chapter 1 : Introduction

1.1 Motivation

Criminal events can be characterized in terms of their location as well as the time at which they occurred. Likewise, other features of criminal events provide a rich source of information from which to draw inferences and conclusions about patterns of crime. It has been shown that certain geographic features such as population density and distance from landmarks can be predictive of the likelihood of future criminal events. There is also evidence that attributes that are a function of the time at which crimes occur, such as the daily temperature or the proximity to a sporting event, are strong predictive attributes. Use of these types of additional information about criminal events can produce significant enhancements to crime analysis tools, and eventually could lead to decreased crime rates.

The primary motivation for this project is to enhance the analytic forecasting capabilities of an existing system, the Regional Crime Analysis Program (ReCAP) developed at the University of Virginia. The ReCAP system is an interactive shared information and decision support system that includes a database, a geographical information system (GIS), and statistical tools. These components are integrated together into a single system to provide the maximum utility and power to users.

While ReCAP currently has tools that perform certain kinds of forecasting based solely on crime event histories, there is a need for an analytic tool that incorporates additional event feature information. We have developed a new tool for ReCAP that not only generates maps that highlight the areas where crimes are most likely to occur, but also identifies key features associated with the data that offer intuitive understanding of crime initiation decisions. In addition, this tool allows crime analysts to propose event features and test whether they are associated with the observed crime patterns. Such a tool provides tremendous advantages over

many of the analysis techniques and visualization methods currently employed by crime analysts, such as identifying crime patterns by examining static "pin maps".

1.2 Objectives

A methodology for generating crime predictions was recently developed by Hua Liu (1999) at the University of Virginia. The fundamental idea supporting this methodology is that criminals make rational decisions as to where and when to commit crimes, and these decisions are guided by preferences for situations exhibiting particular characteristics, such as low risk of arrest or potential for high reward. Given a set of possible characteristics or features, Liu's model attempts to identify a subset of the features that are most strongly correlated with crime incidents in a historical data set and discover the pattern of preferences for each of these features. These inferences are then used to estimate the likelihood of another incident occurring within a geographic region and within a specified time range. In Liu's model, the observed data – the times, locations, and features of space-time events – are viewed as a realization of a marked space-time shock process, and the space-time prediction problem as the estimation of the transition density of the process. This model allows for the inclusion of event-related features into the transition density estimation.

There are three principal objectives of the research described in this paper. The first goal is to incorporate additional types of event features into Liu's existing model for density estimation, especially features in the temporal realm. We also aim to provide an effective forecasting tool and decision-support system for the ReCAP system based on this methodology. A further objective is to carry out testing of the methodology through application to additional sets of data, to include various crime types and study regions.

Examples of temporal features for an event include the time of day or the weather at the time of occurrence and the time elapsed since the occurrence of an external event, such as a sporting event. While the values of spatial features remain constant over time at a particular

geographic location, the values of temporal features are dependent on the time when they are measured. Certain types of temporal features must be modeled differently from other event features, and a model for density estimation that incorporates these features is presented.

The remainder of this paper presents the methodological developments and implementation work that comprise this project, as well as results and evaluations of the work. The second chapter surveys the literature describing research that the present work draws upon. The third chapter formally states the problem being undertaken. The fourth chapter presents the existing approach for accomplishing a space-time event prediction, discusses the changes we have made to extend the methodology, and illustrates the special characteristics of temporal features and the methods developed for handling these features in more detail. The fifth chapter describes the implementation of the model as a tool for use within ReCAP, details the methods employed in testing and evaluating the model, and presents results of testing the model on real-world data. The final chapter includes a summary of this paper, states the contributions made, and proposes additional opportunities for research based on this work.

Chapter 2 : Literature Review

The work presented in this paper draws from the domains encompassing the components of the model, including stochastic point processes, spatial-temporal event prediction, and density estimation. This work also relies on criminological theories that provide insight into crime analysis and motivate the problem. A significant portion of this chapter is dedicated to the dissertation of Hua Liu (1999), as much of the theoretical work presented in this thesis is based on this dissertation. Care is taken throughout this paper to explicitly indicate which work is original research and which work is attributed to Mr. Liu.

2.1 Stochastic Point Processes

Space-time data sets consist of a group of observations taken at specified locations over a range of time. These data sets require special models and methods of analysis to determine the underlying processes that produce the structure of the data.

A stochastic process $\{X(\mathbf{t}), \mathbf{t} \in \mathbf{T}\}$ is a collection of random variables, where for each $\mathbf{t} \in \mathbf{T}$, $X(\mathbf{t})$ is a random variable. In the case of a space-time stochastic process, $\mathbf{t}$ is a vector index of the time and space of the process. In other words, a space-time stochastic process is a family of random variables that describes the evolution through time and space of some process. The idea of a stochastic point process is described in many stochastic modeling texts, such as Ross (1993).

Fiksel (1984) developed two space-time cluster models for predicting the positions of future earthquakes in a region, based on the locations and times of previous earthquakes in that region. One model, the *order-model*, only considers the temporal order of events, while the *instant-model* also considers the relative times of the events. Both models assume a stochastic event process, with spatial and temporal interactions, and the model parameters are estimated using the maximum likelihood procedure.

2.2 Spatial-Temporal Event Prediction

The general class of space-time models handles the case in which the system being modeled exhibits systematic dependence between the observations at each region and the observations at neighboring regions. Just as the correlation measure assesses the level of dependence within a univariate time series, this dependence in a space-time series is called *spatial correlation.*

Most work in the area of spatial-temporal event prediction is related to the STARMA family of models, which are an extension of ARMA univariate time series models into the spatial domain. These models have been shown to be effective in describing the historical patterns of spatial data, when the spatial locations of the data contribute to the observed correlation structure. In the STARMA class of models, observations of the random variable at a particular location are expressed as weighted linear combinations of past observations and errors that may be lagged both in time and space. An extension to the STARMA models was developed by Pfeifer and Deutsch (1980) to incorporate temporal differences, and the resulting models were called STARIMA models. The problem with all of the above models, as we are concerned, is that none of them exploit extra feature information associated with the locations and times of events to aid in the future predictions.

Perhaps the definitive works in this field are the recent publications by Liu (1999) and Brown (1999). The dissertation by Liu forms the basis for this thesis. The primary unique aspect of this work is that the proposed model integrates event characteristics or features into the space-time event prediction.

In these papers, the observed data – the times, locations, and features of space-time events - are viewed as a realization of a marked space-time shock process, and the space-time prediction problem as the estimation of the transition density of the process. This model allows for the inclusion of event-related features into the transition density estimation.

2.3 Density Estimation

In the proposed model by Liu and in this thesis, the heart of the modeling effort in space-time prediction is the estimation of the transition density of the spatio-temporal point process. There are two main components of the estimation procedure in the model – analysis of the data to discover patterns in the feature data and combining information about these clusters to create a single model of the multivariate transition density. In Liu's model, the technique used to analyze patterns in the feature data is *hierarchical clustering*, discussed in Anderberg (1973) and Hartigan (1975).

One of the models employed for components of the transition density is the class of *finite mixture distributions*, using Gaussian distributions for the components, and the Expectation-Maximization (EM) algorithm developed by Dempster, Laird, and Rubin (1977) to obtain maximum likelihood estimates of the Gaussian parameters. Another model used is the class of *kernel density estimators*, first introduced by Rosenblatt (1965) and Parzen (1962), and more specifically the class of *filtered kernel estimators* (Marchette et al., 1996), which use a small number of bandwidth matrices, where each matrix is associated with individual region.

2.4 Criminology

The premise of the *Rational Criminal Theory* is that there are underlying reasons why criminals choose to commit a crime at a particular time in a particular location. It is likely that each criminal has a set of preferences that are taken into account when deciding where and when to execute a crime. These preferences reflect an intention to minimize the risk of being caught while maximizing the expected payoff of committing an offense. This theory is one of the most fundamental for our work, because our model is based on the assumption that there are regular patterns in crime history data. If criminals behave randomly instead of rationally, there is no way that accurate predictions of future criminal activity can be made.

Crime patterns can be studied over time to elucidate regularities in crime occurrences and to investigate the effect of societal changes, including legal, political, demographic, and economic structures. The idea that temporal patterns of social behavior are determinants of the level of crime was first promoted by Cohen and Felson (1979). Their argument is that most crime takes place with the combination of motivated offenders, suitable targets, and the absence of capable guardians. These situations arise based on patterns of "routine activities" in a community, such as daily work schedules or regular weekly outings that produce conditions that are conducive to crime occurrences.

Field (1992) provides an example of the effect of a temporal feature on crime occurrence, in his study of the effect of temperature on crime. This study demonstrated that higher temperatures are positively correlated with crime occurrence rates for many types of crime, while sunshine and rainfall levels do not seem to have an influence.

Brantingham and Brantingham (1984) provide an excellent discussion of spatial patterns of crime, criminal decision-making related to choosing spatial targets, and a survey of a wide range of related literature. In any city, there will be certain paths that are more commonly traveled and certain areas that are characterized by a higher rate of activity. Patterns of crime often match highway paths and aggregate activity spaces.

Duffala (1976) found that convenience stores that were closest to major roads, but not on them, and that had no nearby businesses open at night, had the highest victimization rates. The closeness to major roads made those locations more accessible, while their location slightly away from high traffic areas and activity centers reduced the risk of witnesses or interference.

Several studies have been performed in the last half-century to investigate relationships between spatial patterns of crime and demographic or socioeconomic factors. Lander (1954) and Morris (1958) analyzed delinquency patterns in cities using data on a set of socioeconomic variables as well as land-use data. They both found that the socioeconomic variables could explain most of the variance in delinquency rates across regions.

Chapter 3 : Problem Statement

The primary motivation for this project is to enhance the forecasting capabilities of the Regional Crime Analysis Program (ReCAP), and this inspires the following problem: Given observed data for the locations, times, and feature values for a set of events of the same type, we would like to create a density map representing the likelihood that a crime will occur at each point in the region during a certain time interval in the future. This basic problem is the same as the one encountered by Liu (1999), and as such we will use his notation and setup below, with our own modifications as necessary.

The locations and times of the events $(\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2), \ldots, (\mathbf{s}_n, t_n), t_0 = 0 < t_1 < t_2 < \ldots < t_n$ and their associated features or marks $\mathbf{x}_{s_1,t_1}, \mathbf{x}_{s_2,t_2}, \ldots, \mathbf{x}_{s_n,t_n}$ are viewed as a realization of a *marked space-time shock process* of the form

$$\left\{\mathbf{x}_{\mathbf{s},t} \in \chi : \mathbf{s} \in D, t \in T\right\}$$

where t, $\mathbf{s}$, and $\mathbf{x}_{\mathbf{s},t}$ are all random, D is the geographic study region, and T is the study horizon. Viewing the process as shocked is a minor simplification in our model, as many crime events either take place over time or are virtually instantaneous but the exact time is not known.

Events are located within a study region or geographic space $D \subset \Re^2$ and are indicated by a pair of coordinates, such as longitudes and latitudes, $\mathbf{s} = (s_1, s_2)$. $T \subset \Re^+$ is the range of times when events could occur, and is named the *study horizon*. $\chi \subset \Re^p$ is the collection of possible values of the feature vectors (marks) with p dimensions, and is termed the *feature space*.

In this work, the space-time prediction problem is formulated as the estimation of the *transition density* of the stochastic point process described above. The transition density of the process is formally defined as the probability that a single event occurs within a specified minute region and within a specified minute time interval. Mathematically, given the observed data sets

$T_n = \{t_1, t_2, \ldots, t_n\}$, $D_n = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, and $\chi_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ up to instant t_n, we would like to estimate, for a specific region $\mathbf{s}_{n+1} \in D$ and time $t_{n+1} > t_n$, the transition density as:

$$\psi_n(\mathbf{s}_{n+1}, t_{n+1} \mid D_n, T_n, \chi_n) \equiv \lim_{v(ds_{n+1}), dt_{n+1} \to 0} \frac{\Pr\{N(ds_{n+1}, dt_{n+1}) = 1 \mid D_n, T_n, \chi_n\}}{v(ds_{n+1})dt_{n+1}}$$

where $\mathbf{s}_{n+1}$ and t_{n+1} are realizations of the location and time of the next event, respectively, N counts the number of events in an infinitesimal region ds_{n+1} around $\mathbf{s}_{n+1}$ during an infinitesimal time interval dt_{n+1} around t_{n+1}, and v is the volume of region ds_{n+1}.

3.1 Event Features

The features that define the dimensions of feature space can reflect a wide variety of characteristics of crime scenarios. *Geographic* features include distances from event locations to types of geographic landmarks and demographic characteristics of neighborhoods or districts, such as median household income, population density, and ethnic population distributions. *Temporal* features include the time of day of an event, the day of week, and the amount of time from another event, such as a school closing or a baseball game.

At another level, features can be categorized by the nature of their possible values. Some features have *numerical* values that are defined over all of $\Re$ while some may only have positive values. Certain features are *categorical*, and are limited to a prescribed number of possible values. Other features are considered to be *temporal* and have values corresponding to a time on a 24-hour scale.

Liu classifies the set of features into two groups – those that are *inherently temporal (IT)* and all others. Liu defines "IT features" as those features that "label" time intervals so that categorization of time instants can be obtained. An IT feature such as "season of the year" partitions the time axis into time intervals, where the value of the IT feature is constant for all points in a time interval. If the data is restricted to a single such time interval, the assumption of temporal stationarity is maintained. However, restricting the data in this way excludes the use of

temporal information about the crimes that may have trends or patterns that would improve the accuracy of the predictions.

We chose to relax Assumption 3.1 in Liu's model, that the set of features F only contains features that depend on geographic location and excludes temporal features. It is clear that the occurrence of crimes is not only correlated to spatial features, but also to temporal features, and to interactions between temporal and spatial features. For most types of crimes, there is a general temporal pattern of occurrence.

The interaction of spatial and temporal features is evident in the pattern of auto thefts. During working hours, most cars are parked near offices, while in the evenings and at night, cars are more scattered in residential areas, or gathered near shopping centers or entertainment areas. The association of these patterns with crime incidents has been demonstrated by Boggs (1964) and Mayhew et al. (1976).

It is important to remember that the temporal range of training data sets is limited to no more than a few weeks, so there should not be long-term trends in the data due to the changing of seasons or other long-term effects. We are not concerned with the actual values of the transition density, only the pattern of densities over the prediction region and over a daily or weekly time horizon.

It should be clear that the inclusion of certain types of temporal features in the model does not significantly alter the structure of the model. The reason is that the temporal features we consider are actually *features* of the time - not the absolute times themselves. These features do not affect the stationarity of the model because they are not subject to trends. For example, the temporal feature that is the time-of-day of an event is merely a characteristic of the event, but is not considered in the context of the date of the event. We do not expect the average value of such a feature to exhibit a moving trend over the limited range of the training or prediction horizon. When we calculate the transition density for a future scenario, we consider the spatial characteristics and features of that scenario as well as the temporal features.

Chapter 4 : Methodology

The model originally proposed by Liu attempts to capture the underlying processes driving event occurrences over the study region and the study horizon. Development of the model involves a two-step decomposition of the transition density into components that incorporate various aspects of the modeling approach.

4.1 Model Development

The first step of the decomposition separates the spatial transition and temporal transition. The model becomes

$$\psi_n(s_{n+1}, t_{n+1} \mid D_n, T_n, \chi_n) = \psi_n^{(1)}(s_{n+1} \mid D_n, \chi_n, T_n, t_{n+1}) \cdot \psi_n^{(2)}(t_{n+1} \mid T_n)$$

where the first term, $\psi^{(1)}{}_n(s_{n+1} \mid D_n, \chi_n, T_n, t_{n+1})$, is called the *spatial transition density*, and the second term, $\psi^{(2)}{}_n(t_{n+1} \mid T_n)$, is called the *temporal transition density*. The spatial transition density reflects the probability that an event will occur at location $\mathbf{s}_n$ at any time in the future, given the complete history of prior events in time and space. Similarly, the temporal transition density represents the likelihood that an event will occur at time t_{n+1} in the future, without regard to location or feature values. Liu assumes that the temporal transition density is not dependent on D_n and χ_n, as it would be in a standard Bayesian decomposition.

There is a further decomposition of the spatial transition density that produces components that account more directly for the behavioral theories that we hypothesize are guiding criminal activity. The assumption underlying the model is that criminals select the sites where they commit crimes and the times when they commit the crimes rationally, based on characteristics associated with the locations and times. We will henceforth describe the site location and time of a crime event as a *scenario*.

In describing the decisions as rational, we mean that the offenders consistently demonstrate a preference for scenarios that have similar values of particular attributes or features, and in doing so, reject sites with different feature values. The set of preferences is manifested as a small-variation distribution of values in feature space, or a cluster of values. By studying past event information, we can infer sets of selection preferences and use those to forecast where and when future crimes will occur.

Two assumptions are critical to application of these ideas to the model, when the model is put to use on real data. There will invariably be more than one group of offenders operating in a particular region over a particular time interval, each with a different set of selection preferences. These groups must be considered simultaneously in order to generate a complete prediction of future activity over the region. We must first assume that all offender groups base their selection decisions on a common set of features. Furthermore, we must assume that the set of features we consider in the model matches the set of features considered by the offender groups – the actual features considered. These features are discovered by Liu's model in an initial feature selection process.

Another property of rational scenario selection considers the spatial interaction or dependence between selected sites in the study region. A common belief in the criminology literature, exemplified by the "journey to crime" theory, is that event initiators will choose a geographically closer location to execute the next event, all other factors being equal. As a result, the influence of past events on the prediction at a particular site should be inversely related to the distance from the site. In addition, since for certain types of crimes, event initiators often do not wait long before acting again, the model should diminish the impact of older events on the prediction of future events.

The model for spatial transition density is based on the assumptions about offenders and features discussed above, as well as the properties of rational scenario selection. There is a

component that incorporates patterns in feature space, a component that incorporates patterns in geographic space, and a component that adjusts for prior information about feature distributions.

The formal description of the model for spatial transition density supposes that the set χ_n of feature vectors is partitioned into C disjoint subsets $\{\chi_n^{(j)} : j = 1, 2, \ldots, C\}$ corresponding to the clusters of preferences in feature space. The sets of locations D_n and times T_n are also partitioned into C disjoint subsets. If $\mathbf{x}_{n+1}$ is the set of estimated feature values at location $\mathbf{s}_{n+1}$ and instant t_{n+1}, the model for spatial transition density is

$$\psi_n^{(1)}(s_{n+1} \mid D_n, T_n, \chi_n, t_{n+1}) = \alpha \cdot \psi^{(11)}(\mathbf{x}_{n+1} \mid \chi_n) \cdot \sum_{j=1}^{C} \psi^{(12)}(\mathbf{s}_{n+1} \mid D_n^{(j)}, T_n^{(j)}, t_{n+1}) \Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\}$$

The *first-order spatial transition density*, $\psi^{(11)}(\mathbf{x}_{n+1} \mid \chi_n)$, is the probability that an event will occur at a location with a particular feature vector, based solely on the history of feature vectors. This probability does not depend at all on the location of the future point or the historical points, but rather on the feature values of all previous crimes that occurred at any location. This component captures inferences about the reasons why a criminal chose a particular site, rather than the geographic positioning of that location.

Each *second-order spatial transition density*, $\psi^{(12)}(\mathbf{s}_{n+1} \mid D_n^{(j)}, T_n^{(j)}, t_{n+1})$, indicates the closeness in geographic space of the future point s_{n+1} to all of the past events in a particular cluster. Since there is no way to assign a future point to a cluster with total certainty, this density must be included for all possible clusters. The overall density is calculated as a weighted sum of the second-order transition densities for all clusters. The density for each cluster j is weighted by a *spatial interaction probability*, $\Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}\}$, which reflects the likelihood that the feature vector for the future point, χ_{n+1}, is similar to the feature vector $\chi_n^{(j)}$ that defines the cluster j.

A final adjustment to the model is made to incorporate *a priori* knowledge of the distribution of feature values over the study region D. There is an additional term included in the

model that is necessary when the predicted feature values over the study region are not uniformly distributed.

4.1.1 First Order Spatial Transition Density

Liu chooses two types of estimators to compute the first order spatial transition density. *Finite mixture distributions* are parametric models, while *filtered kernel density estimators* are non-parametric models. The use of both of these estimators in our model requires that the number of distinct local (covariance) structures is known, and if this information is not known *a priori*, it must be estimated from the data using *hierarchical clustering* methods. The number of distinct local structures corresponds to the number of offender groups believed to be represented by the data set. Hierarchical clustering provides a means of discovering natural groupings of observations within the data when the number of groups is not known beforehand.

4.1.1.1 Finite Mixture Distributions

One of the methods used in Liu's model to estimate the first order spatial transition density is the class of finite mixture distributions. Finite mixture distributions are estimations of a density function created from the superposition of multiple component distributions. A finite mixture probability density is the weighted sum of all component densities, where each component density is a function of a vector of variables and a set of parameters, and has the form

$$f(\mathbf{x};\, \pi,\, \Theta) = \sum_{j=1}^{C} \pi_j f_j\left(\mathbf{x};\boldsymbol{\theta}_j\right)$$

where $\pi_j > 0$, $j=1,2,\ldots, C$, $\pi_1 + \pi_2 + \ldots + \pi_C = 1$, $\pi = [\pi_1\ \pi_2\ \ldots\ \pi_C]'$, $\Theta = [\boldsymbol{\theta}_1\ \boldsymbol{\theta}_2\ \ldots\ \boldsymbol{\theta}_C]$. Each $f_j\left(\mathbf{x};\boldsymbol{\theta}_j\right)$ is a *component density* with parameters $\boldsymbol{\theta}_j$ and each π_j is a *mixing weight*. As in the rest of the model, $\mathbf{x}$ is a vector of features. Liu chooses to use Gaussian distributions for the components, and the Expectation-Maximization (EM) algorithm, a numeric method, to obtain maximum likelihood estimates of the Gaussian parameters. However, when there are categorical variables, latent class models are used for component densities.

4.1.1.2 Kernel Density Estimators

Kernel density estimators are another method available in Liu's model to estimate the first order spatial transition density. The class of *filtered product kernel* (FPK) estimators in the model place an individual probability density function or kernel over each cluster in the data set, each with its own bandwidth matrix, and sum the individual density functions over the entire region to form a surface. FPK estimators are a special case of filtered kernel estimators in which the bandwidth matrix **H** is a diagonal matrix. These estimators are given as

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C}\frac{\rho_j(\mathbf{x}_j)}{h_{j1}\cdots h_{jp}}\left\{\prod_{l=1}^{p}K\left(\frac{[\mathbf{x}]_l - [\mathbf{x}_i]_l}{h_{jl}}\right)\right\}$$

where n is the number of instances, p is the number of dimensions of the feature vector, h_{jl} is a local bandwidth for the lth dimension of $[\mathbf{x}]_l$ of cluster j, K is a kernel function, and $\rho_j(\mathbf{x})$ is a *filtering function.*

These estimators include filtering functions that are used to incorporate prior information about the clustered regions. The standard FPK estimator, described by Marchette (1996), suggests the use of a finite mixture model to formulate the filtering functions for each cluster. Another variation constructs these functions without the use of finite mixture models, and these estimators are termed *weighted product kernel* (WPK) estimators.

4.1.2 Second Order Spatial Transition Density

The spatial transition models developed by Fiksel (1984) that were introduced earlier are used to estimate the second order spatial transition densities for each cluster. These models are particularly applicable to our model because they incorporate the "journey to event" theory described in Chapter 2 by giving more influence to events that occurred geographically closer to the future event point of estimation.

For each cluster of past events, the second-order spatial transition density is a function of the distance from each past event in the cluster to the future event location and the times of the

past events in that cluster. Fiksel's order model, which only considers the temporal ordering of the event times, not the actual times, postulates the following function for the density of cluster j when there are m events in that cluster

$$\psi_n^{(12)}(\mathbf{s} \mid D_n^{(j)}, T_n^{(j)}, t) = \varphi_m(\mathbf{s} \mid \mathbf{s}_1, \ldots, \mathbf{s}_m) = \frac{\lambda^2}{2\pi m} \sum_{i=1}^{m} e^{-\lambda \|\mathbf{s}-\mathbf{s}_i\|}$$

where $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$ are ordered by the event times $t_1, t_2, \ldots, t_m$ corresponding to the earliest through the latest event. The likelihood of observing these past m events is given as

$$L(\mathbf{s}_1, \ldots, \mathbf{s}_m; \lambda) = \varphi_{m-1}(\mathbf{s}_m \mid \mathbf{s}_1, \ldots, \mathbf{s}_{m-1}) \cdots \varphi_1(\mathbf{s}_2 \mid \mathbf{s}_1) = \prod_{i=1}^{m-1} \frac{\lambda^2}{2\pi i} \sum_{k=1}^{i} e^{-\lambda \|\mathbf{s}_{i+1}-\mathbf{s}_k\|}$$

The maximum likelihood estimate $\hat{\lambda}$ is obtained by maximizing this equation. Fiksel's instant-model has a similar form, though it also utilizes the actual values of the event times in the model.

4.1.3 Spatial Interaction Probability

There is a spatial interaction probability computed for each cluster, and these elements are defined on the basis of the first-order model utilized, whether a finite mixture distribution or a filtered kernel estimator. In both cases, the spatial interaction probability for a future event and a particular cluster is the proportional contribution of the component transition density for that cluster to the overall transition density of the future event.

4.1.4 Temporal Transition Density

Since this term is defined to depend only on the past event times, it will be a constant value for all locations in the study region. We are primarily concerned with the pattern of the spatio-temporal transition densities over the region, and not the actual magnitudes of the densities, so this constant factor does not need to be considered. This is only the case when the times range over a short range of time in which feature values do not exhibit a trend. Over a longer horizon, techniques such as ARIMA could be used to estimate this density.

4.2 Temporal Feature Analysis

Temporal patterns are unmistakably evident in the occurrences of many types of crime. These patterns can be inferred intuitively as well as through investigation of historical crime data. Bank holdups are much more likely to occur during daytime banking hours, since there would be no one to hold up when the bank is closed. Similarly, one would expect public intoxication offenses to occur most commonly at night, with peaks of activity soon after bar closings or club events. Excluding temporal information from a crime prediction methodology severely limits the ability of the model to accurately forecast crimes and protect against them. We attempt to capture this information by incorporating *time-of-day* features, *event-related* features, and other temporal features.

When investigating patterns in time, the primary consideration is the scale of time that is used, which has been termed the *temporal cone of resolution.* Long-term trends in crime rates can be analyzed over periods of years or decades, while seasonal patterns can be examined over a series of days, weeks, months, or years. The choice of temporal resolution is critical to drawing out the type of information most necessary for the purposes of a particular model.

Though our model generates predictions based on data from multiple weeks, the level of temporal feature resolution that we deem most relevant to our model is on a relatively small scale, over a period of hours or days. The features that we investigate are things like time-of-day, day-of-week, temperature, and the times from particular events. Since all of these features are considered over a relatively short span of time – no more than one or two weeks – it is assumed that the feature patterns do not exhibit any trends, and thus remain stationary. For this reason, these temporal features can be included in the set of features used to compute the first order spatial transition density and the spatial interaction probabilities.

This approach was deemed more effective in capturing temporal patterns than a time series approach. One of the difficulties in using a time series approach in our instance is that the

event data is not measured at regular intervals, which adds complexity to the analysis. Furthermore, since we are most interested in discovering the temporal peaks in the data, cluster analysis can identify the same patterns as can a time series analysis of seasonality. The variety of features that profile the temporal data, such as the time-of-day and day-of-week features, offer the same information as identifying seasonalities at multiple periods.

4.2.1 Time-of-Day Feature

One of the most interesting problems encountered in the development of the new model presented here was the treatment of temporal features with values that exhibit cyclical properties, specifically the "time-of-day" feature. In our model, the "time-of-day" feature looks at the time, from 0:00 to 24:00, without regard to the date or day of week. Unlike most numerical features, such as the distance between two points, or the median income of a census block, there is no concept of a "greater" or "lesser" time, but rather the concept of "before" and "after". More importantly, there is no concept of a minimum or maximum time of day, as the numerical representation of a time of day is merely a reference, not an actual value. For example, 8 AM can be dually considered before or after 6 PM, depending on the point of initial reference.

The unique properties of cyclical features first come into play in the process of grouping the crime data points into clusters. To determine the distance between feature vectors of two data points, the distance between each pair of values for a particular feature must be calculated. The "distance" between a time of 11:00 PM (23:00) and 3:00 AM (03:00) can either be four hours or 20 hours (23:00 – 3:00) depending on the point of reference. The minimum distance would be four hours, even though this distance crosses over the 0:00 line. This minimum distance between two times will be referred to as the *temporal distance*. The issue is further complicated when the mean time-of-day needs to be computed, such as when the cluster centroid must be determined.

Given two times, there are two possible mean values separated by 12 hours. The mean of 11 PM and 3 AM could either be 1 AM or 1 PM, depending the direction traveled around the 24-

hour clock. The mean that is most intuitive is the point halfway during the span representing the temporal distance between the points. In other words, the mean should lie within the most acute angle created between the two "time vectors" created on a 24-hour clock. In the above case, that mean would be 1 AM, since the temporal distance between the two points is four hours, and 1 AM is within that four-hour span. However, when there are more than two times to be averaged, the true value of the mean is not immediately clear.

We chose to identify the true mean as the time that minimizes the sum of squared error (SSE) of the mean and all other times. This criterion makes the most intuitive sense, even though in some cases there will be multiple equally valid mean values by the SSE criterion. The strength of this mean value can be suggested by looking at the standard deviation of times around the mean, and only placing significance on those mean values that are accompanied by a relatively small standard deviation.

We considered a variety of approaches to solving this problem, including a vector-based method, an optimization method, and a method of straight addition. The final method used is a hybrid of a selection of these methods. Any approach must take into account the cyclical properties of the time-of-day feature values.

The *optimization method* calculates the mean of the set of times by minimizing the sum of squared errors (SSE) from the mean time. The errors are calculated as the temporal distances from the mean time from each other time point. This method will find the global minimum SSE, but often only finds local minima when performed by Excel.

The *vector-based method* is rooted in the cyclical nature of the times. The initial step is to map the times onto points on the circumference of a 24-hour clock, and create vectors of unit length from the origin to each of those points. The vector sum of all of the vectors is computed, resulting in a single vector. The angle of this resulting vector is calculated, and this is converted back to a time. This vector-based method results in the exact mean of a set of points in simple

cases, but more often is merely a close approximation of the actual mean. The advantage of this method is that it generates a consistent result no matter the order of inputs.

The third approach considered, the *straight addition method*, corresponds most closely to the method of calculating the mean of a standard set of numerical data. Once a mean has been determined for a set of points, if the mean needs to be adjusted to include an additional point, a weighted average is calculated. The original mean is weighted by the number of points included, and a weighted average is computed with the new point given a weight of one. The problem with the straight addition method is that it is highly dependent on the order of points considered in the mean calculation; different means can be calculated when the order of times is changed.

The method we developed that fulfills all of the problem constraints and can be executed with a small amount of calculation is a hybrid of the vector-based method and the straight addition method. The first step is to calculate the vector-based mean, which is a close approximation of the true mean. Once this approximate mean has been obtained, the straight addition method can be performed using the vector mean as an initial basis. The method assumes that the mean of any two points will lie within the angle formed by the two points and including the vector mean. This method is guaranteed to generate a mean that is equal to the true mean, independent of the ordering of the points. This method also works when the times have weights attached to them, as will often be the case in the procedure for density estimation.

4.2.2 Event-Related Features

Event-related features are those features that measure the difference in time between a crime occurrence and a particular event or type of event. These features are especially adept at addressing the *routine activity theory* discussed in Chapter 2. An example of an event-related feature with a particular event would be the difference in time from the end of a baseball game. There is often a surge in crime surrounding sporting events with a large number of fans in attendance, as emotions and intoxication levels can run high by the time fans leave the stadium.

The value of this feature for each crime event would be the difference between the time of the particular event and the time of the crime occurrence. This feature value could be positive or negative, depending on which event happened earlier, and could span multiple days.

Another event-related feature could be included to measure the difference in time from *any* baseball game to a crime occurrence. The value of this feature for each crime event would be the time to the *nearest* event in time, whether that event is earlier or later. For example, if a crime occurred two hours after the end of the first game of a homestand, but 18 hours before the end of the second game, the value of that feature would be two hours. Features relating crime to an event type could also consider regular events such as daily school closings or bank openings.

Event-related features do not exhibit cyclical properties, since their values are not limited to the 24-hour day, the 7-day week, or any other cycle. In fact, these features are treated in exactly the same manner as regular numerical features.

4.2.3 Other Temporal Features

Another temporal feature is the "day of week" feature, which is a categorical feature. While the day of week displays some characteristics of a cyclical value, values of categorical features are considered to have no hierarchical order in our model. However, it is not clear that any information is lost due to this simplification, because we generally consider days to be separate entities. We would only expect a higher incidence of crime on a Sunday if there had been many crimes on all other Sundays, not because of crime patterns on Saturdays.

Another type of temporal feature is an *event-independent feature.* Values of these features change over time, but are not associated with any particular event. An example of such a feature is the temperature at the time of a crime occurrence, which has been shown by Field (1992) to be correlated with some types of crime. A feature could also be included that represents the level of police deployment at the crime site at the time of occurrence.

Chapter 5 : Application with Results

We have discussed the motivation for development of the theoretical basis for the prediction model. An equally significant aspect of this thesis was the actual implementation of this theory in a working application. The implementation work was motivated by two main goals. The first was to incorporate the theoretical extensions related to temporality into the existing prediction application; the second goal was to prepare the application for integration into a future version of the ReCAP system.

The existing program was transformed from a stand-alone prediction application into a complete decision-support system to aid crime analysts. This new system is named STADIUM, as it is a Spatio-temporal Transition Density Model. STADIUM takes the user from an initial set of basic crime data points through the predictive model-building process, and produces a form of output that provides the user with a useful and intuitive visualization of the results.

5.1 Decision-Support System

The original intention for the research was to provide a crime prediction tool to serve as a decision-support system for crime analysts using the ReCAP system. Decision-support systems assist managers and analysts in making decisions and evaluating the possible consequences of decisions before they are made. STADIUM was designed to aid crime analysts in projecting areas of high-frequency crime activity and discovering the factors underlying crime activity patterns. With this tool, analysts can recommend an efficient allocation of police resources. Additionally, the model can discover predictive features of crime, while analysts can use the tool to test their own hypotheses about the reasons for crime.

Thematic maps are the means of visualizing model output. The model calculates a value of the transition density at each point of a grid overlaying the prediction region, representing the probability that a crime will happen at that location during a particular time interval. Since the

transition density values at different points in the region often differ by many orders of magnitude, it is more informative to display the ordering of points – which has the highest value, which has the next highest value, etc. – rather than the actual values. Each point on the grid is shaded based on the ordering, with darkest shades representing the points with the highest transition density values. In this manner, the pattern of high- and low-density areas can be easily observed. Figure 5.3 below is an example of a thematic map.

Figure 5.3 : Example Thematic Map

5.2 Evaluation Methods

The new model was evaluated using a modified version of percentile scores to compare performance with a basic comparison model. The model used as a basis of comparison is structured in the same way as Liu's model, except that no feature information is included in the model building process or the transition density generation. The only data used to build the model are the geographic locations of the crime points in the training data set. Hypothesis tests were carried out to determine if there is a significant difference between the results of the two models. Also, the performance of the new model with temporal features was compared against

the model without temporal features, to investigate the effect of temporal features on model results.

5.2.1 Percentile Scores

In the original model, no temporal features were allowed to be included in the model, and therefore, the time dimension was effectively removed from the problem. In our model, however, we permit the inclusion of temporal features, and this difference forces changes in the methods for evaluating the model.

The primary evaluation statistics used by Liu are percentile scores, which are approximations of the relative magnitudes of the density estimates. To calculate percentile scores, the density estimate of a test point is first compared against the density estimates of all grid points. The resulting percentile score reflects the percentage of grid point density estimates that are less than the density estimate of the test point.

When temporal features are included in the model, the output is extended in multiple dimensions, as the density estimates change over space and time, instead of just over space. As a result, there is a new grid of density estimates corresponding to each set of future temporal feature values. The percentile scores must compare test points to the "grid slice" of estimates corresponding to the temporal feature vector of the test point. The percentile score, $p_{s,\mathbf{x}_t}$ at location s and feature vector $\mathbf{x}$ at time t is defined by :

$$p_{s,\mathbf{x}_t} = (100/N)\sum_{i=1}^{N} \mathbf{1}\left\{d_{s,\mathbf{x}_t} \geq d_{s_i^g,\mathbf{x}_t}\right\}$$

where $d_{s,\mathbf{x}_t}$ is the density of the test point at location $\mathbf{s}$ with temporal feature vector $\mathbf{x}_t$ at time t, $d_{s_i^g,\mathbf{x}_t}$ is the density estimate of the grid point at that same location with the same feature vector and time, and $\mathbf{1}\left\{d_{s,\mathbf{x}_t} \geq d_{s_i^g,\mathbf{x}_t}\right\}$ is 1 if $d_{s,\mathbf{x}_t} \geq d_{s_i^g,\mathbf{x}_t}$ and 0 otherwise.

5.2.2 Hypothesis Testing

We use two statistical hypothesis tests to determine whether one model outperforms another in a statistically significant way. The first test does a straight comparison of all percentile scores for the two models, and determines whether the new model outperforms the comparison model for significantly more than half of the test incidences. The second test looks at the actual differences between percentile scores in the two models, to determine if the mean difference is significant. This test is valuable because while the first test might indicate that a significant fraction of percentile scores are larger in one model than the other, it could happen that the actual differences between the percentile scores of the two models are relatively small.

5.3 Model Calibration

5.3.1 Data Sets

The data used in the testing of the model was collected from a variety of sources. The historical crime data was drawn from the ReCAP system for Richmond, and this data originally came from the Richmond police departments. Each data set used in testing contains crime events from one- or two-week periods in November and December of 1997. The data range was intentionally limited to no more than two months, so that there would not be significant trends in the data due to seasonal or other changes. This restriction allows us to assume that the temporal transition density remains constant. The demographic feature data used to assign feature values to crime event points and grid points was taken from 1997 census data. The data grid is a regular grid of 2517 points placed over the Richmond area.

5.3.2 Feature Selection

The first step of Liu's methodology is to select a subset of features from a large set of possible features. The subset of features includes those features that best account for the underlying pattern of event occurrences. Unlike some methods of data mining that will only

improve performance given additional information, including more feature information in our model has the potential to diminish prediction accuracy. This behavior occurs because feature vectors are always considered as a whole, and there is no mechanism for weighting a particular dimension of the feature vector more strongly than other dimensions. Of course, a feature that is very highly correlated within a cluster, indicated by a low standard deviation, will have a more pronounced effect on predictions. But if the values of a particular feature do not fit a Gaussian model, then the mixture models will not capture the true nature of the data.

However, for the purposes of demonstration, we have bypassed the feature selection process, with varying results. As a result, in some instances, there is no distinct temporal correlation among the data, so the standard deviations of temporal features are very large. In these cases, the predictions generated using temporal features do not perform significantly better than the predictions with only non-temporal features, and in some cases they even perform significantly worse. In practice, it is likely that the feature selection process would not select temporal features that would produce inferior results.

The non-temporal features used in these tests are "families per unit area", "personal care expenditures per household", and "distance to nearest highway". Preliminary analysis and visual inspection of distributions of these features indicates that the intensity of "Breaking and Entering" incidents is proportional to family density, while most criminal incidents are concentrated in middle-class and poor regions of Richmond. The distance to highway feature has the most pronounced relationship, as criminal incidents are usually very close to highways in Richmond, which is a pattern in accordance with some of the studies mentioned in Chapter 2.

5.3.3 Modeling Parameters

Our goal in testing is not to evaluate and compare the performance of different transition density estimation methods, as this was the focus of the analysis in Liu's work. We are more concerned with evaluating the effects of temporal features on prediction accuracy. As a result,

we chose to only use *finite mixture models* for the first order spatial transition density, as those performed the best in Liu's testing on future data.

5.4 Evaluation Results

A series of tests were performed on data sets for two primary crime types – "breaking and entering" crimes (Codes 501 – 506) and "auto theft" crimes (Codes 701-706). Results of three different models were compared for each data set. The three models were our model including temporal features, Liu's basic model with only non-temporal features, and a comparison model with no features. In each case, three sets of tests were performed for each model. One test studied results when the original two-week training period was resubstituted as the test data, another test studied results with a test set of one week, and the third test studied results with a bi-weekly test set.

Our intention in these tests is not to comprehensively prove that models with temporal features included are inherently more predictive than models without temporal features. We will demonstrate instances where the model was used with certain temporal features as example cases, some of which illustrate how these features can improve model performance. However, it is important to keep in mind that using temporal features is not a solution in all cases – these features should be considered tools to be applied when appropriate. The feature selection process is the best method of determining when temporal features should be included in the analysis.

5.4.1 Breaking and Entering Data

The three training data sets were November 3-16, November 10-23, and November 17-30, with test sets extending through December 14. The tests performed with this data included use of the "time-of-day" feature. We can see in Figures A.1 – A.6 of Appendix A that the model with temporal features outperformed the basic feature model at a 0.05 level of significance for at least one of the hypothesis tests about one-third of the time, most notably on the November 10-23

training data. The model with temporal features also outperformed the comparison model at a 0.10 level of significance for at least one of the hypothesis tests for five of six test sets on the November 3-16 and November 17-30 training data. However, the comparison model outperformed the model with temporal features for two of the three test sets on the November 10-23 training data. For all but one test, the model with temporal features compares with the other models at a smaller p-value for the bi-weekly predictions than the weekly predictions, indicating that time of day is a relatively better predictor over a longer time horizon.

The figures showing the prediction levels over the Richmond study region indicate that the "distance to highway" feature is a major component of the feature-based models (Figures A.7 and A.8) by inspection of the high intensity patterns. These patterns clearly match the paths of major highways in Richmond. The comparison model (Figure A.9) generates a much smoother prediction surface.

5.4.2 Auto Theft Data

The two training data sets of auto theft crimes were November 3-16 and November 17-30, with test sets extending through December 14. The criminal incident counts were relatively stable over this time period (except for December 1-7). For the most part, these tests demonstrated mixed results for the model with temporal features, as shown in Figures B.1 – B.4 of Appendix B. However, on the November 17-30 training data, the temporal model outperformed the comparison model at a 0.05 significance level on the second hypothesis test for all three test sets. The temporal model was significantly outperformed by the basic feature model in the first hypothesis test on the November 3-16 training data, though this is a surprising result given that the raw data (Figure B.1) indicates that the two models had similar means.

Chapter 6 : Conclusion

6.1 Summary

In the course of this work, we have made the following primary contributions :

- Incorporation of additional types of event features into Liu's existing model for density estimation, especially features in the temporal realm
- Development of temporal distance measures and methods for cyclical analysis
- Creation of an effective forecasting tool and decision-support system based on this methodology for future integration into the ReCAP system
- Implementation of an improved user interface, automated tools, and an enhanced visualization of result data sets
- Testing of the methodology through application to additional sets of data, including various crime types

With more types of features available for use in the model, crime analysts will have a greater variety of tools at their disposal with which to predict likely crime scenarios. Analysts can use existing temporal features or propose others that they believe might explain patterns in the data.

The STADIUM program is a dramatically improved means of testing the transition density model and using it in a real-world context. The decision-support system guides users with a range of skills through the entire model-building process, and provides a rich visualization of the prediction results that will be much more useful for crime analysts.

While the use of temporal features was not shown to produce significant improvements in predictions over the previous model in all cases, there are some instances in which temporal features can enhance the prediction with useful information. The feature selection process should be used to identify when temporal features can provide additional insight into criminal event initiation decisions.

6.2 Future Research

There is definitely more research that can be performed to investigate the appropriate use of temporal features in this spatio-temporal transition density model. The first step that should be taken is to modify the feature selection program to include temporal features. The program must also be improved so that it can be integrated into STADIUM.

The model estimation procedures could be changed to incorporate similarity rankings for categorical variables. These changes would be necessary in the calculation of the first-order spatial transition density, and might eliminate the need for latent class finite mixture models.

Another area of potential future work is to further test the model with additional types of temporal features, and to improve the automatic feature data set creation program to include these features. Currently, only certain types of temporal features can be automatically generated, and all others must be manually created.

References

Abramson, L. (1982). On bandwidth variation in kernel estimates: a square root law. *The Annals of Statistics*, **10**, 1217-1223.

Anderberg, M. (1973). *Cluster Analysis for Applications*. Academic Press, New York.

Boggs, S.L. (1964). Urban Crime Patterns. *American Sociological Review*, **30**, 899-908.

Box, G. and Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco.

Brantingham P. & Brantingham P. (1984). *Patterns in Crime.* Macmillan Publishing Company, New York.

Breiman, L., Meisel, W., and Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics*, **19**, 135-144.

Brockwell, Peter J. & Davis, Richard A.(1996). *Introduction to Time Series and Forecasting.* Springer-Verlag, Inc, New York.

Brown, D. (1999). "Predictive Models for Law Enforcement," NIJ Predictive Modeling Cluster Conference, March 8, Washington, D.C.

Cliff, A. and Ord, J. (1973). *Spatial Autocorrelation.* Pioneer, London.

Cohen, L.E., Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, **44**, 588-608.

Curtis, L. (1974). *Criminal Violence: National Patterns and Behavior.* Lexington Books, Lexington, Mass.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society, B*, **39**, 1-38.

Duffala, D.C. (1976). Convenience Stores, Armed Robbery, and Physical Environmental Features. *American Behavioral Scientist*, **20**, 227-246.

Everitt, B. (1984). *An Introduction to Latent Variable Models.* Chapman and Hall, London.

Field, S. (1992). The Effect of Temperature on Crime. *British Journal of Criminology*, **32**(3), 340-351.

Fiksel, T. (1984). Simple Spatial-Temporal Models for Sequences of Geological Events. *Journal of Information Processing and Cybernetics*, **20**, 480-487.

Forgy, E. (1965). Cluster analysis of multivariate data : efficiency vs interpretability of classifications. *Biometrics*, **21**, 768-769.

Gorr, W., Olligschlaeger, A. (1999). "Crime Hot Spot Forecasting: Modeling and Comparative Analysis," NIJ Predictive Modeling Cluster Conference, March 8, Washington, D.C.

Hartigan, J. (1975). *Clustering Algorithms*. Wiley, New York.

Harvey, D. (1977). *Social Justice and the City*. Edward Arnold, London.

Herbert, D. (1977). An Areal and Ecological Analysis of Delinquency Residence: Cardiff 1966 and 1971. *Tijdschrift Voor Economic En Social Geografie*, **68**, 83-99.

Jancey, R. (1966). Multidimensional group analysis. *Australian Journal of Botany*, **14**, 127-130.

Jarvis, G. (1972). "The Ecological Analysis of Juvenile Delinquency in a Canadian City." In Boydell, et al. *Deviant Behavior and Societal Reaction.* Holt, Rinehart, and Winston, Toronto, pp. 195-211.

Kelly, W. (1999). "A GIS Analysis of the Relationship Between Public Order and More Serious Crime," NIJ Predictive Modeling Cluster Conference, March 8, Washington, D.C.

Lander, B. (1954). *Towards an Understanding of Juvenile Delinquency.* Columbia University Press, New York.

Liu, H., Brown, D (1998). *Spatial-Temporal Event Prediction: A New Model.* Department of Systems Engineering, University of Virginia.

Liu, H. (1999). *Space-Time Point Process Modeling: Feature Selection and Transition Density Estimation.* Department of Systems Engineering, University of Virginia.

Marchette, D., Priebe, C., Rogers, G., and Solka, J. (1996). Filtered kernel density estimation. *Computational Statistics,* **11**, 95-112.

Mayhew, P., Clarke, R.V.G., Sturman, R., and Hough, J.M. (1976). *Crime as Opportunity.* Home Office Research Study No. 34. London: HMSO.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, Inc., New York.

Morris, T. (1958). *The Criminal Area: A Study in Social Ecology*. Routledge and Kegan Paul, London.

Neyman, J. (1939) On a new class of "contagious" distributions, applicable in entomology and bacteriology. *Annals of Mathematical Statistics*, **10**, 35-37.

Parzen, E. (1962). On the estimation of a probability density and mode. *The Annals of Mathematical Statistics*, **33**, 1065-1076.

Pfeifer, Phillip E., & Deutsch, Stuart J. (1980). A Three-Stage Iterative Procedure for Space-Time Modeling. *Technometrics,* **22**, 35-47.

Rengert, G. (1975). "Journey to Crime: An Empirical Analysis of Spatially Constrained Female Mobility." Paper read at Association of American Geographers annual meeting, Milwaukee, 1975.

Ridley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, **27**, 832-837.

Ross, S. (1993). *Introduction to Probability Models*. Academic Press, Inc., San Diego.

Student (1907). "On the error of counting with a haemacytometer. *Biometrika*, **5**, 351-360.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Tufte, E. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.

Tufte, E. (1990). *Envisioning Information*. Graphics Press, Cheshire, Connecticut.

Tufte, E. (1997). *Visual Explanations*. Graphics Press, Cheshire, Connecticut.

Appendix A : Breaking and Entering Data Results

Training Set : 11/3 - 11/16 Breaking & Enterings (141 Incidents)						
	With Time of Day		Without Time of Day		Comparison	
Test Set	Mean	St Dev	Mean	St Dev	Mean	St Dev
Resubstitution (11/3 - 11/16)	76.8471	21.4113	79.2235	20.2853	74.5422	20.4387
Weekly (11/17 - 11/23)	70.3547	21.7386	69.5905	22.4512	67.3278	21.8478
Bi-Weekly (11/17 - 11/30)	73.237	21.3703	71.2381	23.3865	67.96	22.3108

Figure A.1 : Basic statistics : November 3-16 B&E data

Training Set : 11/3 - 11/16 Breaking & Enterings (141 Incidents)								
Time of Day Model vs. No Time of Day Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	141	0.382979	-2.7791	0.9973	-2.37646	13.30665	-2.12067	0.983
Weekly	64	0.484375	-0.25	0.5987	0.764179	10.98766	0.556391	0.2877
Bi-Weekly	141	0.496454	-0.08422	0.5319	1.99889	13.84769	1.714041	0.0436
Time of Day Model vs. Comparison Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	141	0.574468	1.768519	0.0384	2.304894	21.86519	1.25172	0.1056
Weekly	64	0.53125	0.5	0.3085	3.026917	22.86928	1.058859	0.1446
Bi-Weekly	141	0.609929	2.610671	0.0045	5.277024	20.59762	3.042157	0.0012

Figure A.2 : Hypothesis test results : November 3-16 B&E data

Training Set : 11/10 - 11/23 Breaking & Enterings (139 Incidents)						
	With Time of Day		Without Time of Day		Comparison	
Test Set	Mean	St Dev	Mean	St Dev	Mean	St Dev
Resubstitution (11/10 - 11/23)	75.2326	21.6106	74.5932	21.7314	80.3783	15.7248
Weekly (11/24 - 11/30)	74.318	23.3344	74.3645	23.6495	76.0837	22.203
Bi-Weekly (11/24 - 12/7)	74.557	24.406	74.5268	24.2802	73.7741	23.3323

Figure A.3 : Basic statistics : November 10-23 B&E data

Training Set : 11/10 - 11/23 Breaking & Enterings (139 Incidents)								
Time of Day Model vs. No Time of Day Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	139	0.61871	2.79902	0.0026	0.63939	3.93454	1.91594	0.0274
Weekly	77	0.55844	1.02565	0.1515	-0.04644	3.39516	-0.12002	0.5478
Bi-Weekly	137	0.56934	1.62328	0.0526	0.03016	3.80938	0.09267	0.4641
Time of Day Model vs. Comparison Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	139	0.38129	-2.79902	0.9974	-5.14573	20.1838	-3.00575	0.9987
Weekly	77	0.46753	-0.5698	0.7157	-1.76566	20.6332	-0.7509	0.7734
Bi-Weekly	137	0.53285	0.76892	0.2206	0.81287	20.7382	0.45878	0.3228

Figure A.4 : Hypothesis test results : November 10-23 B&E data

Training Set : 11/17 - 11/30 Breaking & Enterings (141 Incidents)						
	With Time of Day		Without Time of Day		Comparison	
Test Set	Mean	St Dev	Mean	St Dev	Mean	St Dev
Resubstitution (11/17 - 11/30)	76.9956	20.5829	76.7228	21.159	71.8713	21.9078
Weekly (12/1 - 12/7)	74.2822	26.3142	74.2471	25.302	72.548	24.0762
Bi-Weekly (12/1 - 12/14)	74.3661	24.8114	73.7116	24.7821	70.2903	23.0574

Figure A.5 : Basic statistics : November 17-30 B&E data

Training Set : 11/17 - 11/30 Breaking & Enterings (141 Incidents)								
Time of Day Model vs. No Time of Day Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	141	0.57447	1.76852	0.0392	0.27276	11.3625	0.28504	0.3859
Weekly	60	0.56667	1.0328	0.1515	0.03509	7.53148	0.03609	0.484
Bi-Weekly	131	0.53435	0.78633	0.2148	0.65448	8.38651	0.8932	0.1867
Time of Day Model vs. Comparison Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	141	0.61702	2.7791	0.0027	5.1243	18.4363	3.30043	0.0005
Weekly	60	0.58333	1.29099	0.0985	1.73421	19.942	0.67361	0.2514
Bi-Weekly	131	0.61069	2.53374	0.0057	4.0758	20.5026	2.27531	0.0116

Figure A.6 : Hypothesis test results : November 17-30 B&E data

Figure A.7 : Thematic map of temporal feature model prediction with November 3-16 B&E training data

Figure A.8 : Thematic map of basic feature model prediction with November 3-16 B&E training data

Figure A.9 : Thematic map of comparison model prediction with November 3-16 B&E training data

Appendix B: Auto Theft Data Results

Training Set : 11/3 - 11/16 Auto Thefts (165 Incidents)						
	With Time of Day		Without Time of Day		Comparison	
Test Set	Mean	St Dev	Mean	St Dev	Mean	St Dev
Resubstitution (11/3 - 11/16)	79.9403	18.5618	79.8129	18.5525	82.4112	15.1234
Weekly (11/17 - 11/23)	74.8153	19.8627	75.021	19.0469	72.2726	20.8655
Bi-Weekly (11/17 - 11/30)	74.4772	19.2061	74.9478	18.2969	73.8849	20.0351

Figure B.1 : Basic statistics : November 3-16 Auto theft data

Training Set : 11/3 - 11/16 Auto Thefts (165 Incidents)								
Time of Day Model vs. No Time of Day Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	165	0.397163	-2.64193	0.9959	0.150748	3.035052	0.63801	0.2611
Weekly	79	0.367089	-2.36268	0.9909	-0.20569	3.293455	-0.5551	0.7123
Bi-Weekly	174	0.304598	-5.15507	> 0.9998	-0.47059	3.569151	-1.73922	0.9591
Time of Day Model vs. Comparison Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	165	0.432624	-1.73092	0.9582	-2.7222	15.11353	-2.31364	0.9896
Weekly	79	0.493671	-0.11251	0.5438	2.54271	16.80613	1.344753	0.0901
Bi-Weekly	174	0.494253	-0.15162	0.5596	0.592294	16.33864	0.478185	0.3156

Figure B.2 : Hypothesis test results : November 3-16 Auto theft data

Training Set : 11/17 - 11/30 Auto Thefts (174 Incidents)						
	With Time of Day		Without Time of Day		Comparison	
Test Set	Mean	St Dev	Mean	St Dev	Mean	St Dev
Resubstitution (11/17 - 11/30)	82.5068	17.068	81.4055	16.3625	72.725	21.1091
Weekly (12/1 - 12/7)	74.7306	20.0006	75.6083	21.9447	69.2177	21.938
Bi-Weekly (12/1 - 12/14)	71.9765	21.7527	71.6636	27.0316	69.1323	23.007

Figure B.3 : Basic statistics : November 17-30 Auto theft data

Training Set : 11/17 - 11/30 Auto Thefts (174 Incidents)								
Time of Day Model vs. No Time of Day Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	174	0.528736	0.758098	0.2236	1.101247	11.3032	1.285163	0.0985
Weekly	67	0.432836	-1.09952	0.8643	-0.87761	13.24119	-0.54252	0.7054
Bi-Weekly	151	0.503311	0.081379	0.4681	0.31284	13.93181	0.275932	0.3897
Time of Day Model vs. Comparison Model								
		Test 1			Test 2			
Test Set	Points	Prob	Z-Stat	p-Value	Mean	St Dev	Z-Stat	p-Value
Resubstitution	174	0.701149	5.306686	< 0.0001	9.78176	20.60551	6.261931	< 0.0001
Weekly	67	0.567164	1.099525	0.1357	5.51296	17.58638	2.565935	0.0051
Bi-Weekly	151	0.516556	0.406894	0.3409	2.844235	21.2863	1.641927	0.0505

Figure B.4 : Hypothesis test results : November 17-30 Auto theft data