Mining Preferences from Spatial-Temporal Data

Donald E. Brown, Hua Liu
and Yifei Xue

Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA 22903
804-924-5393
804-982-2972 (fax)
brown@virginia.edu

Abstract

The discovery of preferences in space and time is important in a variety of applications. In this paper we first establish the correspondence between a set of preferences in space and time and density estimates obtained from observations of spatial-temporal features recorded within large databases. We perform density estimation using both kernel methods and mixture models. The density estimates constitute a probabilistic representation of preferences. We then present a point process transition density model for space-time event prediction that hinges upon the density estimates from the preference discovery process. The added dimension of preference discovery through feature space analysis enables our model to outperform traditional preference modeling approaches. We demonstrate this performance improvement using a criminal incident database from Richmond, Virginia. Criminal incidents are human-initiated events that may be governed by criminal preferences over space and time. We applied our modeling technique to breaking and entering crimes committed in both residential and commercial settings. Our approach effectively recovers the preference structure of the criminals and enables one-week ahead forecasts of threatened areas. This capability to accommodate all measurable features, identify the key features, and quantify their relationship with event occurrence over space and time makes this approach applicable to domains other than law enforcement.

1. Introduction

The concept of evaluating a decision, product, or service as a function of the attributes of alternatives is a rather universally accepted approach, which has been implemented in such fields as economics (McFadden, 1973, 1980; Theil, 1970), transportation (Currim, 1982), finance (Slovic et al., 1972), medicine (Huber et al., 1969), and marketing (Gensch, 1979; Rust and Donthu, 1995). The goal of the research is to analyze the individuals' decision making process and predict the actual choice of particular individuals. Some research, especially in the fields of economics, transportation and marketing have used the analysis of choice behavior, which is first introduced by Luce (1959). The research analyzes and predicts the decision of individuals by their preference on the attributes of alternatives.

Criminal incidents, like many other human-initiated events, are frequently linked with the decision making process and 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. 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

[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.

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).

It is rather safe to say that offenders' preferences constitute an important piece of information to inform future site selection decisions by criminals. 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. 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.

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. We term the geographic region of interest a *study region* or *geographic space* $D \subset \Re^2$, and the time ranges 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. 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*. Suppose that the initial set of features contains those attributes that the event initiators **actually** factor into their decision-making. 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. We can just find the smallest feature subset (key feature set) and the key feature space by feature selection process. 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?

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., $\mathbf{s}_6$ 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 $\mathbf{s}_5$). The merit of integrating feature space information into space-time event prediction is that **potential** event areas can be picked out. The same rationale applies to the analysis of event occurrences in time.

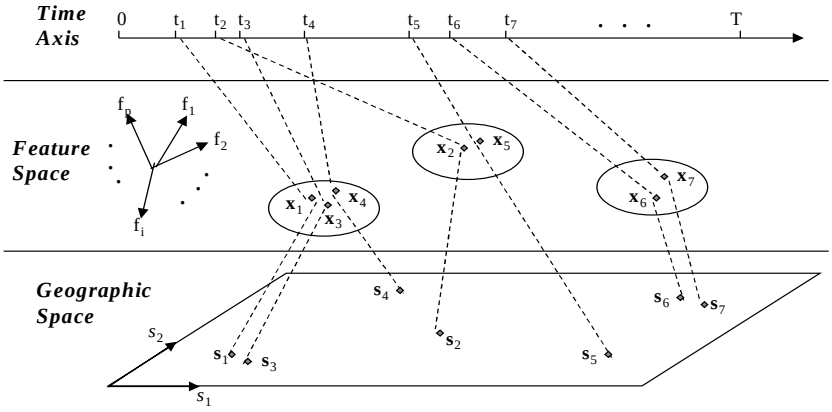

Figure 1. Event occurrences in three hyperspaces.

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.

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

Where $\mathbf{s}_{n+1}$ and t_{n+1} are the location and the time of the next incident, respectively, $\nu\left(d\mathbf{s}_{n+1}\right)$ is the Lebesgue measure of $d\mathbf{s}_{n+1}$ and $N\left(d\mathbf{s}_{n+1}, dt_{n+1}\right)$ 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.

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.

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.

The first step in the process of componentization is to separate spatial and temporal transitions.

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

Where $\psi_n^{(1)}\left(\mathbf{s}_{n+1} \middle| D_n, \chi_n, T_n, t_{n+1}\right)$ will be called *spatial transition density* and $\psi_n^{(2)}\left(t_{n+1} \middle| T_n\right)$ *temporal transition density*. Equation (2) would be a standard Bayesian decomposition if the second term on the right-hand side were $\psi_n^{(2)}\left(t_{n+1} \middle| D_n, \chi_n, T_n\right)$. 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. 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 $\psi_n^{(1)}\left(\mathbf{s}_{n+1} \middle| D_n, \chi_n, T_n, t_{n+1}\right)$. We assume that the features selected initially are the key features. By doing so, we postpone the feature selection task until next subsection. Suppose that the set χ_n of feature vectors is partitioned into C disjoint subsets $\left\{\chi_n^{(j)} : j = 1,2,...,C\right\}$, each of which is mapped onto a clique in key feature space. Corresponding to $\left\{\chi_n^{(j)} : j = 1,2,...,C\right\}$, the set D_n (T_n) of locations (times) of past events is also partitioned into C disjoint subsets $\left\{D_n^{(j)} : j = 1,2,...,C\right\}$ ($\left\{T_n^{(j)} : j = 1,2,...,C\right\}$). 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

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

Where $\psi_n^{(11)}\left(\mathbf{x}_{n+1}\middle|\chi_n\right)$ 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. $\psi_n^{(12)}\left(\mathbf{s}_{n+1}\middle|D_n^{(j)},T_n^{(j)},t_{n+1}\right),\ j=1,2,...,C,$ 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. $\Pr\left\{\mathbf{x}_{n+1}\in\chi_n^{(j)}\right\},\ j=1,2,...,C,$ 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. 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*. 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.

The second order effects are modeled in geographic space. 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).

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 $\kappa_n\left(\mathbf{x}_{n+1}\right)$ denote the probability density function of $\mathbf{x}_{n+1}$ over all predicted feature values for locations $\mathbf{s}_{n+1}\in D$. We adjust (3) as follows.

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

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

We do not give space-time prediction in our case due to the two assumptions we made when we separated spatial and temporal transitions (see Equation (2)). 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.

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*, we have to "discover" it from the data. To accomplish this first task, we use a hierarchical clustering algorithm to generate partitions and

[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.

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. 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) 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

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

Where $\overline{\alpha}_j$ 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).

We consider two classes of models for estimating the first order spatial transition density. The first class is called *finite mixture distributions* (e.g., Everitt and Hand, 1981; Titterington et al., 1985; McLachlan and Basford, 1988). A finite mixture probability density function (or mass function in the case of discrete sample space) has the form

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

Where $\pi_j > 0$, $j = 1,2,...,C$, $\pi_1 + \pi_2 + ... + \pi_C = 1$, $\boldsymbol{\pi} = [\pi_1 \quad ... \quad \pi_C]'$, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \quad ... \quad \boldsymbol{\theta}_C]$. $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ is the *j*th *component density* with the set $\boldsymbol{\theta}_j$ of parameters and $\pi_1, \pi_2, ..., \pi_C$ are *mixing weights*. $\boldsymbol{\Theta}$ 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 $\{\mathbf{x}_i : i = 1,2,..,n\}$.

The component densities $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$ ($j = 1,2,...,C$) are assumed to be fitted by Gaussian mixture models (GMM) for continuous feature space, and fitted by Latent Class Models (LCM) (see Everitt, 1984) for the case of discrete feature space. When mixed variable types are present, it is trivial to combine GMM and LCM provided that the numeric dimensions are independent of the categorical ones. For the set of parameters $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \quad ... \quad \boldsymbol{\theta}_C]$, we use a numeric maximum likelihood algorithm known as Expectation-Maximization (EM) algorithm (see, for example, Dempster, Laird and Rubin, 1977). 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 **x**, are *filtering functions*. We only consider a special case of (7) 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 *l*th dimension $[\mathbf{x}]_l$ of the *j*th 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 *l*th 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 a finite mixture distribution is involved to model first order spatial transition density, 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. Two models developed by Fiksel (1984), known as the order model and the instant model are used. Both models incorporate the "journey to event" assumption (*event initiators are in favor of the geographically closer location to start the next event*). At the same time, the instant model also takes into account the assumption regarding "lingering period to resume act" (*event initiators tend not to wait long before the act again*). 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*

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

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

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

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* is 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.

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.

In this paper we look at a class of cohesiveness measures that do not require any partitioning in advance. These measures are functions of inter-event distances (or similarities). 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

$$s_{ij} = \frac{1}{1+\alpha d_{ij}}, \quad (17)$$

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

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

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.

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

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. 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. 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[3]. Second, 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.

$$\text{Adjusted } I_g = \text{(unadjusted) } I_g \,/\, I_g^p \qquad (20)$$

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". Traditional space-time prediction models do not include feature data and criminal 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.

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. 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.

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

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

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. We assume that the feature values at any given location in the study region remain unchanged within the study horizon.

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. We adjust

[3] 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.

the score with the I_g score obtained based on the feature data pertaining to 2517 locations placed evenly over the Richmond. 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.

Figure 2. B &E criminal incidents between July 1, 1997 and August 31, 1997 in Richmond, Virginia.

Feature	I_g	Adj. I_g		Feature	I_g	Adj. I_g
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 2. 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 3. 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 4. Distance features evaluation result.

We select one feature from each table to form the key feature set so as to avoid strong correlation between any two features in the key feature set. 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.

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 the test scenarios of predicting out one week into the future (weekly 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.

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. The percentile score $p_{\mathbf{s}}$ at location **s** is defined by

$$p_{\mathbf{s}} = (100/N)\sum_{i=1}^{N} \mathbf{1}\left\{d_{\mathbf{s}} \geq d_{\mathbf{s}_i^g}\right\} \tag{21}$$

where $N = 2157$; $\mathbf{s}_i^g$ is the location of the *i*th grid point; $\mathbf{1}\left\{d_{\mathbf{s}} \geq d_{\mathbf{s}_i^g}\right\}$ is 1 if $d_{\mathbf{s}} \geq d_{\mathbf{s}_i^g}$ 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.

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 5, 7, 9, 11, 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 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 $\hat{\delta} > 0.5$; otherwise, test against H_a: $\delta < 0.5$).

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

$$\hat{\delta} = (1/m)\sum_{i=1}^{m} \mathbf{1}\left\{p_{\mathbf{s}_i}^p > p_{\mathbf{s}_i}^c\right\}. \tag{22}$$

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 $\hat{\mu} > 0$; otherwise test against H_a: $\mu < 0$).

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

$$\hat{\mu} = (1/m)\sum_{i=1}^{m} \left(p_{\mathbf{s}_i}^p - p_{\mathbf{s}_i}^c\right). \tag{23}$$

The standard deviation of the difference $q_{\mathbf{s}_i} = p_{\mathbf{s}_i}^p - p_{\mathbf{s}_i}^c$ is estimated by

$$\hat{\sigma} = \left(1/(m-1)\right)\sum_{i=1}^{m} \left(q_{\mathbf{s}_i} - \hat{\mu}\right)^2. \tag{24}$$

The results of these tests are reported in tables 6, 8, 10, and 12, 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 6), 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.

Training set: July 7-20 (145 incidents)				
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	

Table 5. Basic statistics for models calibrated on July 7-20 data.

Training set: July 7-20 (145 incidents)							
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

Table 6. Hypothesis tests results for models calibrated on July 7-20 data.

Training set: July 14-27 (143 incidents)				
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	

Table 7. Basic statistics for models calibrated on July 14-27 data.

Training set: July 14-27 (143 incidents)							
Weekly prediction - Test set: July 28-August 3 (68 incidents).							
	Test 1			Test 2			
Comparison	Prob.	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

Table 8. Hypothesis tests results for models calibrated on July 14-27 data.

Training set: July 21-August 3 (140 incidents)				
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	

Table 9. Basic statistics for models calibrated on July 21-August 3 data.

Training set: July 21-August 3 (140 incidents)							
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

Table 10. Hypothesis tests results for models calibrated on July 21-August 3 data.

Training set: July 28-August 10 (137 incidents)				
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	

Table 11. Basic statistics for models calibrated on July 28-August 10 data.

Training set: July 28-August 10 (137 incidents)							
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

Table 12. 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 (i.e., the test sets) are plotted on the density maps to enable visual examination of how well the proposed model performs under weekly 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. This is confirmed in Tables 5, 7, 9 and 11 where the

GMM version is indeed picked as the "best model" for every weekly 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 (left), WPK (middle) and FPK (right) versions of the proposed model calibrated on July 7-20 data and tested on July 21-27.

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.

Currim, S. Imran. (1982). Predictive testing of consumer choice models not subject to independence of irrelevant alternatives. *Journal of Marketing Research, 208-222.*

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.

Gensch H. D. and Recker W. W. (1979). The multinomial, multiattribute logit choice model. *Journal of Marketing Research*, 124-132.

Huber, George. (1969). Multiplicative utility models in cost effectiveness analyses. *Journal of Industrial Engineering*, 19.

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.

Luce, R. (1959). Individual choice behavior: A theoretical analysis. *New York: John Wiley & Sons, Inc.*

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.

McFadden, Daniel. (1973). Conditional logit analysis of qualitative choice behavior. *Frontiers in Economics*. New Youk: Academic Press.

McFadden, Daniel. (1980). Econometric models for probabilistic choice among products. *Journal of Business*, V53, 13-29.

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.

Rust, T. R. and Donthu, N. (1995) Capturing geographically localized misspecification error in retail store choice models. *Journal of Marketing Research* 103-110

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.

Slovic, P., D. Fleisnner, and W. S. Bauman. (1972). Analyzing the use of information in investment decision making. A methodological proposal. *Journal of Business, April.*

Thiel, Henri. (1969). Amultinomial extension of the linear logit model. *International Economic Review,* 10, 251-259

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