



ELSEVIER

International Journal of Forecasting 19 (2003) 603–622

international journal  
of forecasting

www.elsevier.com/locate/ijforecast

# Criminal incident prediction using a point-pattern-based density model

Hua Liu<sup>a</sup>, Donald E. Brown<sup>\*.b</sup>

<sup>a</sup>CSG Systems, Inc., One Main Street, Cambridge, MA 02142, USA

<sup>b</sup>Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA 22903, USA

---

## Abstract

Law enforcement agencies need crime forecasts to support their tactical operations; namely, predicted crime locations for next week based on data from the previous week. Current practice simply assumes that spatial clusters of crimes or “hot spots” observed in the previous week will persist to the next week. This paper introduces a multivariate prediction model for hot spots that relates the features in an area to the predicted occurrence of crimes through the preference structure of criminals. We use a point-pattern-based transition density model for space–time event prediction that relies on criminal preference discovery as observed in the features chosen for past crimes. The resultant model outperforms the current practices, as demonstrated statistically by an application to breaking and entering incidents in Richmond, VA. © 2003 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

*Keywords:* Point-pattern methods; Crime forecasting; Spatial transition density model; Hot spot prediction model

---

## 1. Introduction

Law enforcement agencies have a continuing need to predict the locations of crimes. Armed with criminal event predictions, police can target patrols, direct surveillance, and conduct other operations to prevent crimes and enforce laws. These activities have horizons of days and, at most, weeks.

It is well known that crimes tend to cluster spatially in so-called hot spots; for example, due to certain crime-prone land uses (e.g., convenience stores or bars) or established patterns of serial criminals. Thus, the prevalent approach to forecasting by police is a simple spatial clustering method using only the coordinates, dates, and types of

crimes. Police assume that current crime clusters will persist over the forecast horizon. A widely used such method is the Spatial and Temporal Analysis of Crime program (STAC) which clusters crime points within ellipses (Block, 1995). Jefferis (1998) surveys additional hotspot methods, the most sophisticated of which employ kernel density estimation (Levine, 1998).

This paper extends crime clustering methods by incorporating offenders’ preferences in crime site selection. A number of researchers have investigated spatial decision making by criminals (Amir, 1971; Baldwin & Bottoms, 1976; Brantingham & Brantingham, 1975, 1984; Capone & Nichols, 1976; LeBeau, 1987; Molumby, 1976; Newman, 1972; Repetto, 1974; Rossmo, 1993, 1996; Scarr, 1973). Taken together, this body of research suggests that the likelihood of a criminal incident at a specified location is based on past incidents of the same type

---

\*Corresponding author. Tel.: +1-434-982-2074.

E-mail addresses: [hua\\_liu@csgsystems.com](mailto:hua_liu@csgsystems.com) (H. Liu), [brown@virginia.edu](mailto:brown@virginia.edu) (D.E. Brown).

and independent spatial attributes or features (e.g. distance to a road, type of residential community, etc.).

To formally describe the forecast problem, we denote the locations and times of criminal incidents as  $(\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2), \dots, t_0 = 0 < t_1 < t_2 < \dots$ , where  $\mathbf{s}_i$  is the two-dimensional location of incident  $i$  of a given crime type and  $t_i$  is the corresponding time. Suppose that there are  $p$  measurable features  $f_1, f_2, \dots, f_p$  that are believed to be relevant to the occurrence of the incidents, with values consisting of  $p$ -dimensional vectors  $\mathbf{x}_1, \mathbf{x}_2, \dots$ . Taken together, the locations, times, and features of all incidents are a realization of a marked space–time shock point process  $\{\mathbf{x}_{s,t} \in \chi: \mathbf{s} \in D, t \in T\}$ , where  $t, \mathbf{s}$ , and  $\mathbf{x}_{s,t}$  are all random quantities confined within a study horizon  $T \subset \mathfrak{R}^+$ , a study region or geographic space  $D \subset \mathfrak{R}^2$ , and a feature space  $\chi \subset \mathfrak{R}^p$ , respectively. The space–time point process is marked by the feature vectors, and is a shock process because the events of the process are considered instantaneous, as opposed to a survival process (Cressie, 1993).

The quantity of interest is the density of the process, which formally captures the likelihood that a future criminal 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. Let  $T_n = \{t_1, t_2, \dots, t_n\}$ ,  $D_n = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$ , and  $\chi_n = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$  where  $s_i = (s_{i1}, s_{i2})$  and  $\mathbf{x}_i = [\mathbf{x}_{i1} \cdots \mathbf{x}_{ip}]'$ . The transition density is defined as follows:

$$\begin{aligned} & \psi_n(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n, \chi_n) \\ & \equiv \lim_{\nu(\mathbf{ds}_{n+1}), dt_{n+1} \rightarrow 0} \frac{\Pr\{N(\mathbf{ds}_{n+1}, dt_{n+1}) = 1 | D_n, T_n, \chi_n\}}{\nu(\mathbf{ds}_{n+1}) dt_{n+1}} \end{aligned} \quad (1)$$

where  $\mathbf{s}_{n+1}$  and  $t_{n+1}$  are the location and the time of the next incident, respectively,  $\nu(\mathbf{ds}_{n+1})$  is the Lebesgue measure of the infinitesimal region  $\mathbf{ds}_{n+1}$  and  $N(\mathbf{ds}_{n+1}, dt_{n+1})$  counts the number of incidents that occur within  $\mathbf{ds}_{n+1}$  and the infinitesimal time interval  $dt_{n+1}$ .

In this paper, we examine and evaluate a model of the transition density that we derive from the theory of point patterns (Diggle, 1983). The model represents criminal preferences as the functional relation-

ships between demographic, economic, social, victim, and spatial attributes and measures of criminal activity. The model represents a declining or rising trend as a decreasing or increasing likelihood of being victimized by crime. Furthermore, the model can identify new potential hot spots or areas at risk that are not necessarily in the vicinity of existing crime locations.

In Section 2, we give an overview of our model specification process, and then discuss a Gini-index-based measure of cohesiveness for feature selection. In Section 3, we present a model of the transition density as our approach for spatio-temporal event prediction. We also describe a comparison model in that section. In Section 4, we calibrate our proposed model on crime data from Richmond, VA, and compare it to the current hot spot approach. Included are two sets of hypothesis tests. We briefly conclude in Section 5.

## 2. Model search method

Our objective is to find the smallest feature subset (of the initial feature set) that accounts for the underlying pattern of criminal event occurrences (hot spots). This is a model search problem we call feature selection. The selected feature subset is called the key feature set and the feature subspace defined by the key feature set the key feature space.

A feature selection problem is specified by a triplet  $(F, c, s)$ , where  $F$  is the initial feature set,  $c$  a criterion function defined for subsets of  $F$ , and  $s$  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 feature ranking and will be the approach that we use for our application in Section 4.

To evaluate a given set of features, we need a measure of cohesiveness of a point pattern observed in the independent variable or feature subspace defined. In this paper we employ a class of cohesiveness measures that do not require any partitioning of space in advance. These measures are functions of inter-event distances (or similarities). Let  $d_{ij}$  be the distance between two events  $i$  and  $j$  in the feature subspace defined by the feature subset to be evalu-

ated. We transform the distance  $d_{ij}$  into the similarity  $s_{ij}$  as follows.

$$s_{ij} = \frac{1}{1 + \alpha d_{ij}} \tag{2}$$

where  $\alpha = 1/\bar{d}$  and  $\bar{d}$  is the average inter-event distance, where distance refers to differences in value of an independent variable. Define the Gini index between these two events as

$$g_{ij} = 4s_{ij}(1 - s_{ij}) \tag{3}$$

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

$$I_g = \frac{2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n g_{ij}}{n(n-1)} \tag{4}$$

The smaller the value of the  $I_g$  index is, the higher the level of point-pattern cohesiveness or the better the set of features that define the point pattern.

In general,  $I_g$  can be used in a subset selection algorithm (e.g., forward selection, backward elimination) to yield an optimal or suboptimal subset of features. Alternatively, one can also evaluate  $I_g$  for each individual feature and select a subset of features based on the  $I_g$  scores. We adopt this latter approach for our application in Section 4. Suppose that in addition to actual event data, we also have the feature values for a large sample of locations that are chosen uniformly over the study region. We call the set of the feature values at the sample locations the prior feature data set. As the first feature selection step, we calculate the ratio of the observed range to the full range of each feature dimension to see whether there are any dimensions that do not exhibit enough variation in the event feature data set. This ratio for feature  $f_k$  is defined by

$$r_k = \frac{\max_{x_{ik}, x_{jk} \in E_k} |x_{ik} - x_{jk}|}{\max_{x_{ik}, x_{jk} \in P_k} |x_{ik} - x_{jk}|} \tag{5}$$

where  $E_k$  and  $P_k$  are the event and the prior feature data sets for feature  $f_k$  (i.e., containing only the dimension  $f_k$ ), respectively. If the ratio  $r_k$  is consid-

ered sufficiently small, we will not calculate the  $I_g$  score for feature  $f_k$ . Otherwise, we calculate the adjusted  $I_g$  for feature  $f_k$ , or the adjusted  $I_g^{(k)}$ , defined as follows.

$$\text{Adjusted } I_g^{(k)} = \frac{I_g(E_k)}{I_g(P_k)} \tag{6}$$

where  $I_g(E_k)$  and  $I_g(P_k)$  are the  $I_g$  scores for feature  $f_k$  over the event feature data set  $E_k$  and the prior feature data set  $P_k$ , respectively. The rationale for this adjustment scheme is that  $I_g(P_k)$  indicates how much the prior distribution of  $f_k$  deviates from the uniform distribution. The smaller  $I_g(P_k)$  is or the further the prior distribution is from the uniform distribution, the more  $I_g(E_k)$  is adjusted.

### 3. The transition density model

The development of our model for space–time prediction involves a multi-step componentization of the transition density (Eq. (1)) and the estimation of corresponding model components. We describe the componentization process in this section but estimation methods in Appendix A. To model the transition density (Eq. (1)), we first decompose it into spatial and temporal components. The spatial component incorporates interactions between times, locations, and features and represents all aspects of site selection behavior. An adjustment factor removes the effect of nonuniform prior distribution of the features on the predicted density. Our model is schematically represented in Fig. 1, and we give details of each component in the sequel.

The first step in the componentization process is to separate spatial and temporal transitions. We postulate that the occurrences of events (criminal incidents) over time and space are separable as follows.

$$\begin{aligned} \psi_n(\mathbf{s}_{n+1}, t_{n+1} | D_n, T_n, \chi_n) \\ = \psi_n^{(1)}(\mathbf{s}_{n+1} | D_n, \chi_n, T_n, t_{n+1}) \cdot \psi_n^{(2)}(t_{n+1} | T_n) \end{aligned} \tag{7}$$

where  $\psi_n^{(1)}(\mathbf{s}_{n+1} | D_n, \chi_n, T_n, t_{n+1})$  is called the spatial transition density and  $\psi_n^{(2)}(t_{n+1} | T_n)$  the temporal transition density. Eq. (7) would be a standard Bayesian decomposition if the second term on the right-hand side were  $\psi_n^{(2)}(t_{n+1} | D_n, \chi_n, T_n)$ . The

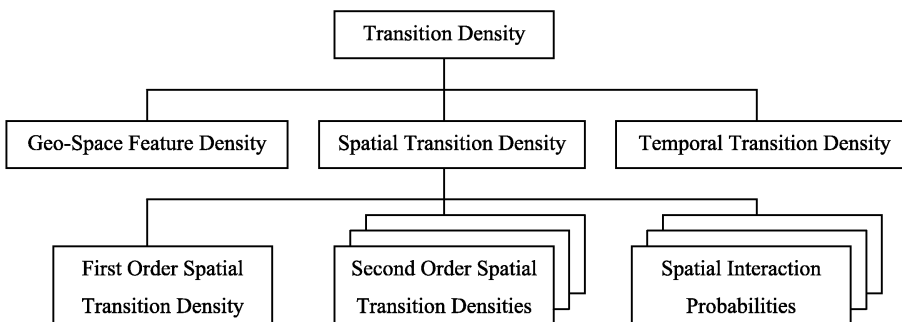


Fig. 1. Components of the transition density model.

feature data set  $\chi_n$  and the location data set  $D_n$  of past events were left out under the following assumptions.

First, we do not need to consider any inherently temporal features (e.g., season of the year and holiday/nonholiday) that categorize time instants. We exclude such features because we deal with short time series, within an estimation period of a week or a few weeks. Also, as is typical for space–time point processes (Cressie, 1993), temporal transition of the marked space–time shock point process is assumed not to depend on its spatial transition. In the criminal event scenario, this assumes that the distribution of past crime locations ( $D_n$ ) does not influence how soon criminals are going to strike again ( $t_{n+1}$ ).

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 preferences to inform how likely future events are to occur at certain locations. We know from the last section that site selection preferences are defined by a distinct clustering pattern in the key feature space. Suppose that the key feature space ( $\chi$ ) is composed of  $C$  disjoint continuums  $\{\chi^{(j)}: j=1, 2, \dots, C\}$  in relation to some underlying clustering pattern which defines the sets of preferences. The set  $\chi_n$  of feature vectors is then partitioned into  $C$  disjoint subsets  $\{\chi_n^{(j)}: j=1, 2, \dots, C\}$  where  $\chi_n^{(j)} \subset \chi^{(j)}$ . Corresponding to  $\{\chi_n^{(j)}: j=1, 2, \dots, C\}$ , the sets  $D_n$  of locations and  $T_n$  of times of past events are also partitioned into  $C$  disjoint subsets  $\{D_n^{(j)}: j=1, 2, \dots, C\}$  and  $\{T_n^{(j)}: j=1, 2, \dots, C\}$ , respectively. Let  $\mathbf{x}_{n+1}$  be the feature vector at location  $\mathbf{s}_{n+1}$

and instant  $t_{n+1}$ . The spatial transition density is assumed to take the form

$$\begin{aligned} &\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1}) \\ &= \alpha \cdot \psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n) \cdot \sum_{j=1}^C \psi_n^{(12)} \\ &\quad \times (\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1}) \Pr\{\mathbf{x}_{n+1} \in \chi^{(j)}|\chi_n^{(j)}\} \end{aligned} \tag{8}$$

where  $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$  is called the first-order spatial transition density which reflects the event intensity (i.e., first-order effects) at  $\mathbf{x}_{n+1}$  in the key feature space;  $\psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1})$ ,  $j=1, 2, \dots, C$ , are called second-order spatial transition densities which describe the interaction (i.e., second-order effects) of a new event location  $\mathbf{s}_{n+1}$  with past event locations in each  $D_n^{(j)}$ , respectively;  $\Pr\{\mathbf{x}_{n+1} \in \chi^{(j)}|\chi_n^{(j)}\}$ ,  $j=1, 2, \dots, C$ , are called spatial interaction probabilities which are the probabilities that  $\mathbf{x}_{n+1}$  falls in the same continuum  $\chi^{(j)}$  of the key feature space as  $\chi_n^{(j)}$  does;  $\alpha$  is a normalizing factor.

Model (Eq. (8)) incorporates all aspects of site selection behavior in a formal framework—the theory of spatial point patterns. 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 event intensity in feature space which reflects the potential of alternative sites to attract future events, rather than event intensity in geographic space which is simply the expected number of accumulated events at alternative sites. The notion of site selection preferences, which is more fitting for prediction given the assumption that

the same preferences will persist to  $t_{n+1}$ , is captured only by feature space event intensity.

We do not consider second-order effects in feature space because we further assume that the point process in the key feature space is a Markovian process of a small range (Cressie, 1993). Broadly speaking, this assumption ensures that in the key feature space, there are no second-order effects (i.e., interaction or dependence) between clusters, and because the range is small, only first-order effects are significant within each cluster. This assumption formally characterizes the point pattern in the key feature space or the site selection behavior revealed by feature space analysis.

We model second-order effects in geographic space. Note that it is only appropriate to examine spatial interaction among events in the same feature space cluster because these events are initiated with the same set of preferences. However, due to the uncertainty associated with assigning a new event to a specific cluster (or claiming that a new event is representative of a specific set of preferences), we weigh second-order effects pertaining to individual clusters by the probabilities that quantify this uncertainty (i.e., the spatial interaction probabilities). Technically, we calculate the weighted average of the second-order effects of  $C$  thinned (or selected) point processes in geographic space. The thinning from the overall process is based on membership in the relevant cluster. A realization of each thinned point process is the set  $D_n^{(j)}$  of events corresponding to those that form the cluster  $\chi_n^{(j)}$  in the key feature space. Additional assumptions on site selection behavior concerning “journey to event” and “lingering period to resume act” may be captured by the models we use for second-order effects (see Appendix A).

The spatial transition density model (Eq. (8)) needs “prior” adjustment when the collection of feature vectors ( $\mathbf{x}'_{n+1}$ s) for uniformly and independently sampled event locations within the study region ( $D$ ) does not form a uniform distribution. Let  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$  denote the probability density function of  $\mathbf{x}_{n+1}$  given a prior probability density function of  $\mathbf{s}_{n+1}$  over the study region  $D$ . Without any past observations, it would be reasonable to assume that the occurrence of events in geographic space follows a homogeneous Poisson point process (or complete

randomness). In other words, an event is equally likely to occur at any location  $\mathbf{s}_{n+1} \in D$  and any two events are independent. Hence event locations will form a uniform distribution over time. However, this property does not necessarily hold true in feature space due to the form of the mapping from  $\mathbf{s}_{n+1}$  to  $\mathbf{x}_{n+1}$  and the possible inherent randomness of  $\mathbf{x}_{n+1}$ . Nonuniformity of  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$  given a uniform distribution of event locations indicates that 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 Eq. (8) as follows.

$$\begin{aligned} \psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1}) &= \beta \cdot (1/\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})) \cdot \psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n) \\ &\cdot \sum_{j=1}^C \psi_n^{(12)}(\mathbf{s}_{n+1}|D_n^{(j)}, T_n^{(j)}, t_{n+1}) \\ &\times \Pr\{\mathbf{x}_{n+1} \in \chi_n^{(j)}|\chi_n^{(j)}\} \end{aligned} \tag{9}$$

where  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$  is called the *geographic-space feature density* and  $\beta$  is a normalizing factor. Note that the fundamental difference between  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$  and  $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$  is that  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$  is a “prior” density because it does not depend on event feature data  $\chi_n$  while  $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$  is a “posterior” density because it does. When  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$  is uniform, there is no “prior” effect to be adjusted out of  $\psi_n^{(11)}(\mathbf{x}_{n+1}|\chi_n)$  and model (9) reduces to Eq. (8). We use Eq. (8) when we do not have knowledge of  $\kappa_n(\mathbf{x}_{n+1}|\mathbf{s}_{n+1})$ .

The Eqs. (7)–(9) collectively define the transition density model that we propose for spatial–temporal event prediction. For our purpose, the estimation of the individual components of the model requires the following four tasks:

- (1) Partition the event feature 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 give the details of the procedures and density estimation models involved in the above four tasks in Appendix A. The density estimation models we use for the individual components distinguish different versions of our model.

The astute reader may ask why we do not need to estimate the temporal transition density. The answer is that we generally do 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 Eq. (7). 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 by 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 factors in Eqs. (8) and (9), respectively.

In the next section, we calibrate several versions of our model on crime data and compare them with counterpart hot spot methods. Unlike our model, hot spot prediction models do not include feature data nor do they extrapolate based on criminal preferences over these feature data. Ignoring the feature data, a hot spot 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  $(\mathbf{s}_1, t_1), (\mathbf{s}_2, t_2), \dots, (\mathbf{s}_n, t_n), t_0 = 0 < t_1 < t_2 < \dots < t_n < t_{n+1}$ . The quantity of interest is the density function  $\psi_n(\mathbf{s}_{n+1}, t_{n+1}|D_n, T_n)$ , where  $T_n = \{t_1, t_2, \dots, t_n\}$  and  $D_n = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n\}$ . Under the assumption that the occurrence of events over time and space are independent, the class of hot spot models is specified by

$$\psi_n(\mathbf{s}_{n+1}, t_{n+1}|D_n, T_n) = \psi_n^{(1)}(\mathbf{s}_{n+1}|D_n) \cdot \psi_n^{(2)}(t_{n+1}|T_n) \quad (10)$$

In parallel with our model, we term  $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n)$  the spatial transition density and  $\psi_n^{(2)}(t_{n+1}|T_n)$  the temporal transition density. However, unlike the spatial transition density  $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n, \chi_n, T_n, t_{n+1})$  in our model,  $\psi_n^{(1)}(\mathbf{s}_{n+1}|D_n)$  only captures event evolution 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. The temporal transition density  $\psi_n^{(2)}(t_{n+1}|T_n)$  is the same quantity as in our model. For the same reasons we have stated for our model, we can ignore  $\psi_n^{(2)}(t_{n+1}|T_n)$  when presenting the results of a hot spot model as density maps or comparing them with those of our model.

#### 4. Model evaluation: an application

In this section, we apply our proposed transition density model to a sample of crime data. We calibrate several versions of our model and their counterpart hot spot models and compare the results statistically of the two classes of models. To be fair in our comparison, we use exactly the same techniques for hot spot clustering as we use in our model. Hence, the only difference is that our model also includes clustering and preference discovery in feature space. As a result, our evaluation with 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.

##### 4.1. The data

Our sample includes 579 commercial and residential “breaking and entering” (B&E) incidents that occurred in Richmond, VA, between July 1, 1997 and August 31, 1997. Table 1 provides 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 unknown, we choose not to use the data from the first week of July and the last 2 weeks of August for model building in the sequel.

Fig. 2 shows the locations of the B&E incidents. The subregions on the map are census block groups.

Table 1  
Weekly counts of breaking and entering criminal incidents between July 1, 1997 and August 31, 1997 in Richmond, VA

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

We consider three categories of features related to B&E incidents. Demographic and consumer expenditure features data are converted from the 1997 estimates of census categories recorded in [CensusCD+maps \(1998\)](#). The distances from crime locations to geographic landmarks are generated by a

geographic information system ([Brown, 1998](#)). The three categories of feature variables are listed in [Tables 2–4](#), respectively. We assume that the feature values at any given location in the study region remain unchanged within the study horizon (July and August of 1997).

#### 4.2. Model specification

The collection of 60 features shown in [Tables 2–4](#) is our initial feature set. To select the key features from this collection, we first calculate the  $I_g$  score, according to Eq. (4), for each initial feature within the set of feature data pertaining to the B&E incidents between July 7, 1997 and July 20, 1997. This gives us the unadjusted  $I_g$  scores,  $I_g(E_k)$ ,  $k = 1, 2, \dots, 60$ . To remove the influence of the prior feature distribution of  $I_g(E_k)$  scores, we need feature data at uniformly and independently sampled loca-

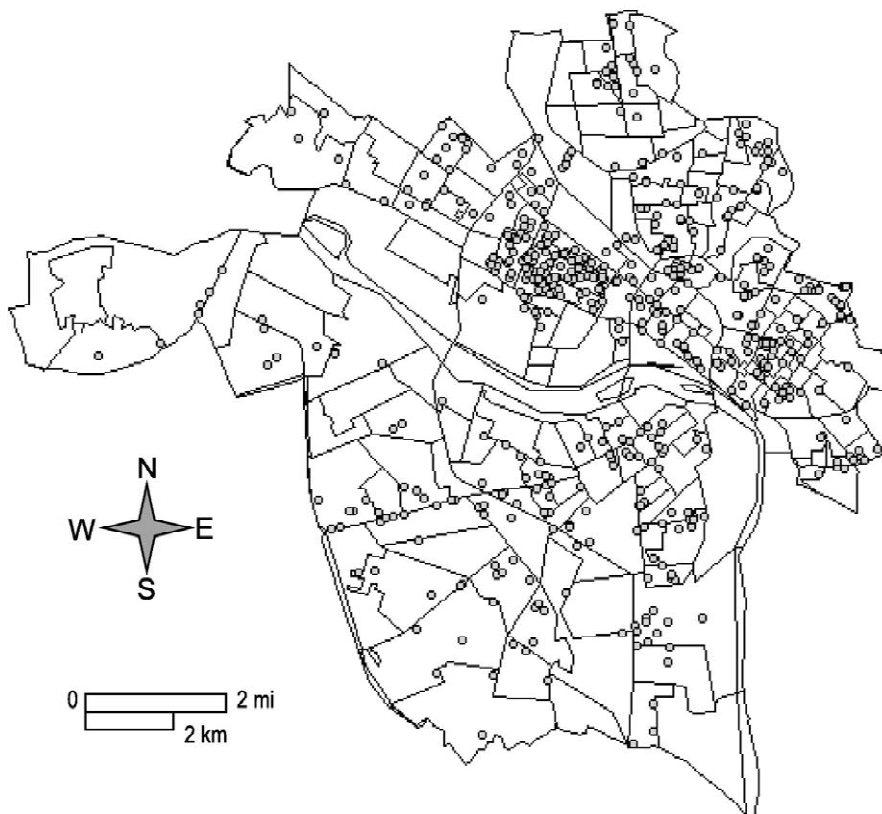


Fig. 2. Point locations of the breaking and entering criminal incidents between July 1, 1997 and August 31, 1997 in Richmond, VA.

Table 2  
Demographic features

Feature	Description
<i>Population, general</i>	
POP_DST	Population per square mile (psm)
HH_DST	Households psm
FAM_DST	Families psm
MALE_DST	Male population psm
FEM_DST	Female population psm
<i>Work force</i>	
CLS12_DST	Private wage and salary workers psm
CLS345_DST	Government workers psm
CLS67_DST	Self-employed and unpaid family workers psm
<i>Income</i>	
PCINC_97	Per capita annual income
MHINC_97	Median annual household income
AHINC_97	Average annual household income
<i>Householder age</i>	
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
<i>Household size</i>	
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
<i>Housing structure</i>	
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
<i>Housing, miscellaneous</i>	
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 3  
Consumer expenditure 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 4  
Distance 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

tions. For our purpose, we place a uniform square grid over the Richmond map and obtain feature values at each grid point. Every two adjacent grid points are separated by  $0.003^\circ$  horizontally and  $0.002^\circ$  vertically. This results in a set of 2517 sampled feature data points, the prior feature data set. We obtain the scores  $I_g(P_k)$ ,  $k=1,2,\dots,60$ ,

using the prior feature data set, and then calculate the adjusted  $I_g^{(k)}$  score for each feature according to Eq. (6). The results are reported in Tables 5–7. Note that before we computed the  $I_g$  scores, we examined the ratio of the observed range to the full range of each feature dimension (according to Eq. (5)) to see whether there are any dimensions that do not exhibit enough variation in the event feature data set. This ratio is greater than 0.2 for every feature in our initial pool. Hence we deem that there is enough variation in every feature for evaluation with the  $I_g$  index and potential inclusion in our model.

We choose one feature from each table to form the key feature set for this study. The features chosen based on adjusted  $I_g^{(k)}$  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 (owner occupied condominiums per square mile) and HSTR9\_DST (occupied trailers per square mile) which have lower adjusted  $I_g^{(k)}$  scores than FAM\_DST. These two features have unusually low  $I_g(P_k)$  scores (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 is sensible because out of the 207 block groups in Richmond there are relatively few 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  $I_g^{(k)}$  scores. Practically, we eliminate these features because we find crime analysts unwilling to accept that the lack of trailer homes or condominiums is linked to higher rate of B&E incidents.

Maps show 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. Secondly, average household expenditure on personal care products and services is an indicator of disposable income within a block group. Most of the B&E 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

Table 5  
Demographic features evaluation result

Feature	$I_g(P_k)$	Adjusted $I_g^{(k)}$
<i>Population, general</i>		
FAM_DST	0.795	0.971
FEM_DST	0.781	1.017
HH_DST	0.766	1.019
POP_DST	0.778	1.022
MALE_DST	0.774	1.038
<i>Work force</i>		
CLS12_DST	0.763	0.996
CLS67_DST	0.718	1.014
CLS345_DST	0.755	1.020
<i>Income</i>		
PCINC_97	0.747	1.094
MHINC_97	0.741	1.101
AHINC_97	0.701	1.169
<i>Householder age</i>		
AGEH12_DST	0.690	0.979
AGEH56_DST	0.759	1.018
AGEH34_DST	0.777	1.048
<i>Household size</i>		
PPH1_DST	0.698	0.999
PPH2_DST	0.774	1.019
PPH3_DST	0.770	1.020
PPH6_DST	0.648	1.096
<i>Housing structure</i>		
HSTR9_DST	0.210	0.430
HSTR6_DST	0.579	0.971
HSTR1_DST	0.780	1.037
HSTR4_DST	0.604	1.096
HSTR10_DST	0.511	1.171
HSTR2_DST	0.514	1.335
HSTR3_DST	0.442	1.543
<i>Housing, miscellaneous</i>		
COND1_DST	0.284	0.250
OCCHU_DST	0.766	1.019
MORT1_DST	0.779	1.034
HUNT_DST	0.765	1.036
OWN_DST	0.780	1.052
RENT_DST	0.691	1.054
OCCHU_PC	0.756	1.070
HUNT_PC	0.762	1.072
MORT2_DST	0.747	1.075
VACHU_DST	0.690	1.088

Table 6  
Consumer expenditure features evaluation result

Feature	$I_g(P_k)$	Adjusted $I_g^{(k)}$
<i>Per household</i>		
P_CARE_PH	0.779	0.887
TRANS_PH	0.748	0.962
MED_PH	0.792	0.970
ET_PH	0.789	0.979
HOUSING_PH	0.697	1.006
REA_PH	0.784	1.016
APPAREL_PH	0.784	1.019
EDU_PH	0.759	1.021
ALC_TOB_PH	0.785	1.025
FOOD_PH	0.749	1.044
<i>Per capita</i>		
P_CARE_PC	0.805	0.958
EDU_PC	0.803	0.979
HOUSING_PC	0.807	0.980
APPAREL_PC	0.814	0.998
ET_PC	0.816	0.999
TRANS_PC	0.821	1.001
ALC_TOB_PC	0.817	1.008
MED_PC	0.813	1.013
FOOD_PC	0.804	1.014
REA_PC	0.799	1.015

Table 7  
Distance features evaluation result

Feature	$I_g(P_k)$	Adjusted $I_g^{(k)}$
D_HIGHWAY	0.803	0.995
D_PARK	0.799	1.004
D_SCHOOL	0.757	1.029
D_CHURCH	0.796	1.033
D_HOSPITAL	0.798	1.036

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.

#### 4.3. Model comparison

We evaluate three versions of our proposed model against their counterpart hot spot models. The three versions are named GMM, WPK, and FPK (see Appendix A). The GMM uses Gaussian mixture models for estimating the first-order spatial transition density, the geographic-space feature density, and 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. All three versions of the model use Fiksel's order model to estimate second-order spatial transition densities, once the model for the first-order spatial transition density is chosen.

We estimate these models and their counterpart hot spot model on four training data sets: B&E incidents that occurred during fortnights, July 7 to 20, July 14 to 27, July 21 to August 3, and July 28 to August 10, respectively. We obtain predictions from each model and compare results for two horizons: weekly and biweekly prediction. Weekly prediction uses the event data from the subsequent week as a holdout sample, while biweekly prediction uses the subsequent 2 weeks. For every version of the proposed model, we use the same set of key features just selected in estimation (e.g., based on the feature data of the incidents between July 7 and July 20), and apply the same prior feature data set (i.e., the 2517 sampled feature data points) to geographic-space feature density estimation.

To compare the results of different models, we convert density estimates into percentile scores which are on a common scale of 0 to 100. Suppose that in addition to the actual crime locations, we have a large set of  $N$  sample locations selected uniformly and independently over the study region. Let  $\mathbf{s}_i^g$  be the  $i$ th sample location or grid point. Denote the density estimate (generated by either a proposed model or the comparison model) at an arbitrary location  $\mathbf{s}$  as  $d_s$ . The predicted percentile score  $p_s$  of location  $\mathbf{s}$  is defined by

$$p_s = (100/N) \sum_{i=1}^N \mathbf{1}\{d_s \geq d_{\mathbf{s}_i^g}\} \quad (11)$$

where  $\mathbf{1}\{d_s \geq d_{\mathbf{s}_i^g}\}$  is 1 if  $d_s \geq d_{\mathbf{s}_i^g}$  and 0 otherwise. Given that the sample set is large enough (or the grid is fine enough) to represent the entire study region, percentile scores are re-scaled density estimates. The higher the percentile score of a specified location is the more likely a new event is to happen at that location.

The model evaluation statistics are mean (predicted) percentile score and sample standard deviation of the mean. For a given holdout set of actual

event locations used for model evaluation, we obtain predictions (density estimates) for all event locations in the set and then convert them into percentile scores. The mean percentile score is the average of these (predicted) percentile scores. We include the two statistics for the three versions of the proposed model and their counterpart hot spot models calibrated on the four aforementioned training data sets in Tables 8–11, respectively. The “best model” in these tables refers to 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.

Two hypothesis tests are performed to confirm these results. Assume that the test data set contains  $m$  incidents that occurred at the locations  $s_1, s_2, \dots, s_m$ , respectively. For the incident at  $s_i$ , let the percentile score given by a proposed model be  $p_{s_i}^p$  and that given by the comparison model be  $p_{s_i}^c$ . Let  $\delta$  be the probability that the proposed model outperforms the

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

Training set: July 7–20 (145 incidents)				
Model type	Proposed model		Comparison model	
	Mean	S.D.	Mean	S.D.
<i>Estimation—Test set: July 7–20 (145 incidents)</i>				
GMM	86.0	15.0	58.3	21.0
WPK	89.5	12.3	83.0	16.9
FPK	89.5	12.3	83.0	16.9
Best model	WPK or FPK		WPK or FPK	
<i>Weekly prediction—Test set: July 21–27 (72 incidents)</i>				
GMM	076.3	26.3	56.5	22.8
WPK	75.9	25.3	74.0	26.6
FPK	75.8	25.3	74.0	26.6
Best model	GMM		WPK or FPK	
<i>Biweekly prediction—Test set: July 21–August 3 (140 incidents)</i>				
GMM	75.9	24.1	57.0	23.1
WPK	74.4	25.2	72.5	26.1
FPK	74.2	25.2	72.5	26.1
Best model	GMM		WPK or FPK	

Table 9  
Basic statistics for models calibrated on July 14–27 data

Training set: July 14–27 (143 incidents)				
Model type	Proposed model		Comparison model	
	Mean	S.D.	Mean	S.D.
<i>Estimation—Test set: July 14–27 (143 incidents)</i>				
GMM	81.1	21.2	61.6	25.0
WPK	85.7	15.8	79.8	20.0
FPK	85.8	15.5	79.8	20.0
Best model	FPK		WPK or FPK	
<i>Weekly prediction—Test set: July 28–August 3 (68 incidents)</i>				
GMM	76.3	21.6	59.3	27.6
WPK	72.6	25.3	70.1	27.6
FPK	72.3	25.3	70.1	27.1
Best model	GMM		WPK or FPK	
<i>Biweekly prediction—Test set: July 28–August 10 (137 incidents)</i>				
GMM	73.6	24.25	57.5	26.5
WPK	72.0	26.5	69.8	27.5
FPK	71.8	26.5	69.8	27.5
Best model	GMM		WPK or FPK	

Table 10  
Basic statistics for models calibrated on July 21–August 3 data

Training set: July 21–August 3 (140 incidents)				
Model type	Proposed model		Comparison model	
	Mean	S.D.	Mean	S.D.
<i>Estimation—Test set: July 21–August 3 (140 incidents)</i>				
GMM	79.78	19.68	60.14	26.31
WPK	80.74	19.09	77.11	21.06
FPK	80.68	18.97	77.11	21.06
Best model	WPK		WPK or FPK	
<i>Weekly prediction—Test set: August 4–10 (69 incidents)</i>				
GMM	73.33	23.88	54.35	25.33
WPK	69.35	28.31	67.26	29.69
FPK	69.28	28.24	67.26	29.69
Best model	GMM		WPK or FPK	
<i>Biweekly prediction—Test set: August 4–17 (141 incidents)</i>				
GMM	77.12	22.53	55.71	25.73
WPK	72.73	27.01	71.66	27.65
FPK	72.56	27.00	71.66	27.65
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)				
Model type	Proposed model		Comparison model	
	Mean	S.D.	Mean	S.D.
<i>Estimation—Test set: July 28–August 10 (137 incidents)</i>				
GMM	79.0	20.4	44.4	26.5
WPK	80.4	19.3	75.6	22.8
FPK	80.4	19.0	75.6	22.8
Best model	FPK		WPK or FPK	
<i>Weekly prediction—Test set: August 11–17 (72 incidents)</i>				
GMM	81.7	20.4	38.5	25.9
WPK	76.2	25.0	75.5	25.0
FPK	76.0	25.0	75.5	25.0
Best model	GMM		WPK or FPK	
<i>Biweekly prediction—Test set: August 11–24 (126 incidents)</i>				
GMM	81.0	20.9	40.4	25.4
WPK	76.7	23.9	75.4	24.0
FPK	76.5	24.0	75.4	24.0
Best model	GMM		WPK or FPK	

comparison model on a single prediction. The null hypothesis for our first hypothesis test is as follows:

$$H_0: \delta = 0.5$$

vs. the alternative

$$H_a: \delta > 0.5$$

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

$$H_a: \delta < 0.5.$$

The test statistic  $\hat{\delta}$  for this hypothesis test is:

$$\hat{\delta} = (1/m) \sum_{i=1}^m \mathbf{1}\{p_{s_i}^p > p_{s_i}^c\} \tag{12}$$

where  $\mathbf{1}\{p_{s_i}^p > p_{s_i}^c\}$  is 1 if  $p_{s_i}^p > p_{s_i}^c$  and 0 otherwise.

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

$$H_0: \mu = 0$$

vs. the alternative

$$H_a: \mu > 0$$

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

$$H_a: \mu < 0.$$

The test statistic  $\hat{\mu}$  based on a test set of  $m$  incidents is straightforward:

$$\hat{\mu} = (1/m) \sum_{i=1}^m (p_{s_i}^p - p_{s_i}^c). \tag{13}$$

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

$$\hat{\sigma} = (1/(m - 1)) \sum_{i=1}^m (q_{s_i} - \hat{\mu})^2 \tag{14}$$

The results of these hypothesis tests are reported in Tables 12–15, in which “probability”, “mean”, and “S.D.” correspond to  $\hat{\delta}$ ,  $\hat{\mu}$ , and  $\hat{\sigma}$ , respectively. The  $z$ -statistic and  $p$ -value are calculated for each hypothesis test to indicate whether the test passes (rejects its null hypothesis in favor of its alternative) or fails (cannot reject its null hypothesis in favor of its alternative) and the significance of the result. These tables indicate that

- for all but one comparison, the proposed 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 12), the performances of the two models are statistically indistinguishable since the two tests are set up against opposite alternative hypotheses but neither test can reject its null hypothesis in favor of its 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.

Density maps generated by the three versions of

Table 12  
Hypothesis tests results for models calibrated on July 7–20 data

Training set: July 7–20 (145 incidents)							
Comparison	Test 1			Test 2			
	Probability	z-Statistic	p-Value	Mean	S.D.	z-Statistic	p-Value
<i>Estimation—Test set: July 7–20 (145 incidents)</i>							
GMM vs. GMM	0.883	9.22	<0.001	27.7	26.3	12.71	<0.001
WPK vs. WPK	0.938	10.55	<0.001	6.5	7.9	9.87	<0.001
FPK vs. FPK	0.910	9.88	<0.001	6.5	8.1	9.69	<0.001
Best vs. Best	0.910	9.88	<0.001	6.5	8.1	9.69	<0.001
<i>Weekly prediction—Test set: July 21–27 (72 incidents)</i>							
GMM vs. GMM	0.750	4.24	<0.001	19.8	32.5	5.17	<0.001
WPK vs. WPK	0.583	1.41	0.079	2.0	11.0	1.53	0.063
FPK vs. FPK	0.597	1.65	0.050	1.8	11.0	1.43	0.076
Best vs. Best	0.444	0.94	0.174	2.3	19.4	1.02	0.154
<i>Biweekly prediction—Test set: July 21–August 3 (140 incidents)</i>							
GMM vs. GMM	0.729	5.41	<0.001	18.9	31.2	7.15	<0.001
WPK vs. WPK	0.586	2.03	0.021	1.86	7.8	2.80	0.003
FPK vs. FPK	0.586	2.03	0.021	1.64	8.0	2.42	0.008
Best vs. Best	0.479	0.51	0.305	3.32	15.8	2.49	0.006

the proposed model built on the training data set of the 145 incidents between July 7 and July 20 are given in Fig. 3. The criminal incidents occurring

within the immediate following week and 2 weeks (i.e., the test sets for weekly and biweekly prediction scenarios) are plotted on the density maps to enable

Table 13  
Hypothesis tests results for models calibrated on July 14–27 data

Training set: July 14–27 (143 incidents)							
Comparison	Test 1			Test 2			
	Probability	z-Statistic	p-Value	Mean	S.D.	z-Statistic	p-Value
<i>Estimation—Test set: July 14–27 (143 incidents)</i>							
GMM vs. GMM	0.783	6.77	<0.001	19.5	28.00	8.31	<0.001
WPK vs. WPK	0.902	9.62	<0.001	5.9	7.54	9.33	<0.001
FPK vs. FPK	0.902	9.62	<0.001	5.9	7.70	9.23	<0.001
Best vs. Best	0.902	9.62	<0.001	5.9	7.70	9.23	<0.001
<i>Weekly prediction—Test set: July 28–August 3 (68 incidents)</i>							
GMM vs. GMM	0.809	5.09	<0.001	17.06	27.71	5.08	<0.001
WPK vs. WPK	0.603	1.70	0.045	2.47	8.35	2.44	0.007
FPK vs. FPK	0.588	1.46	0.072	2.16	8.51	2.09	0.018
Best vs. Best	0.544	0.73	0.233	6.17	14.78	3.44	<0.001
<i>Biweekly prediction—Test set: July 28–August 10 (137 incidents)</i>							
GMM vs. GMM	0.766	6.24	<0.001	16.07	29.57	6.36	<0.001
WPK vs. WPK	0.620	2.82	0.002	2.25	8.70	3.02	0.001
FPK vs. FPK	0.577	1.79	0.037	2.06	8.80	2.74	0.003
Best vs. Best	0.518	0.43	0.334	3.83	16.74	2.68	0.004

Table 14  
Hypothesis tests results for models calibrated on July 21–August 3 data

Training set: July 21–August 3 (140 incidents)							
Comparison	Test 1			Test 2			
	Probability	z-Statistic	p-Value	Mean	S.D.	z-Statistic	p-Value
<i>Estimation—Test set: July 21–August 3 (140 incidents)</i>							
GMM vs. GMM	0.829	7.78	<0.001	19.6	27.1	8.59	<0.001
WPK vs. WPK	0.857	8.45	<0.001	3.6	5.5	7.75	<0.001
FPK vs. FPK	0.857	8.45	<0.001	3.57	5.8	7.24	<0.001
Best vs. Best	0.857	8.45	<0.001	3.6	5.5	7.75	<0.001
<i>Weekly prediction—Test set: August 4–10 (69 incidents)</i>							
GMM vs. GMM	0.797	4.94	<0.001	19.0	29.9	5.28	<0.001
WPK vs. WPK	0.565	1.08	0.140	2.1	10.8	1.60	0.055
FPK vs. FPK	0.580	1.32	0.093	2.0	11.00	1.53	0.063
Best vs. Best	0.580	1.32	0.093	6.1	19.2	2.62	0.004
<i>Biweekly prediction—Test set: August 4–17 (141 incidents)</i>							
GMM vs. GMM	0.830	7.83	<0.001	21.4	28.0	9.08	<0.001
WPK vs. WPK	0.532	0.76	0.224	1.1	4.9	2.60	0.005
FPK vs. FPK	0.553	1.26	0.104	0.9	5.0	2.15	0.016
Best vs. Best	0.560	1.43	0.076	5.5	16.9	3.83	<0.001

visual examination of how well the proposed model performs under weekly and biweekly prediction scenarios, respectively. It is easily seen on these

maps that most of the incidents in the test sets indeed occurred around the predicted high-density areas. Also by visual inspection, the GMM version of the

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

Training set: July 28–August 10 (137 incidents)							
Comparison	Test 1			Test 2			
	Probability	z-Statistic	p-Value	Mean	S.D.	z-Statistic	p-Value
<i>Estimation—Test set: July 28–August 10 (137 incidents)</i>							
GMM vs. GMM	0.839	7.95	<0.001	34.6	38.6	10.49	<0.001
WPK vs. WPK	0.891	9.145	<0.001	4.9	8.4	6.79	<0.001
FPK vs. FPK	0.832	7.78	<0.001	4.9	8.6	6.61	<0.001
Best vs. Best	0.832	7.78	<0.001	4.9	8.6	6.61	<0.001
<i>Weekly prediction—Test set: August 11–17 (72 incidents)</i>							
GMM vs. GMM	0.889	6.60	<0.001	43.1	36.0	10.17	<0.001
WPK vs. WPK	0.597	1.65	0.050	0.8	6.0	1.08	0.140
FPK vs. FPK	0.611	1.89	0.029	0.5	6.0	0.73	0.233
Best vs. Best	0.528	0.47	0.319	6.2	18.0	2.92	0.002
<i>Biweekly prediction—Test set: August 11–24 (126 incidents)</i>							
GMM vs. GMM	0.897	8.91	<0.001	40.5	37.0	12.26	<0.001
WPK vs. WPK	0.611	2.49	0.006	1.4	9.8	1.56	0.059
FPK vs. FPK	0.611	2.49	0.006	1.1	9.8	1.29	0.099
Best vs. Best	0.524	0.54	0.298	5.5	18.8	3.31	0.001

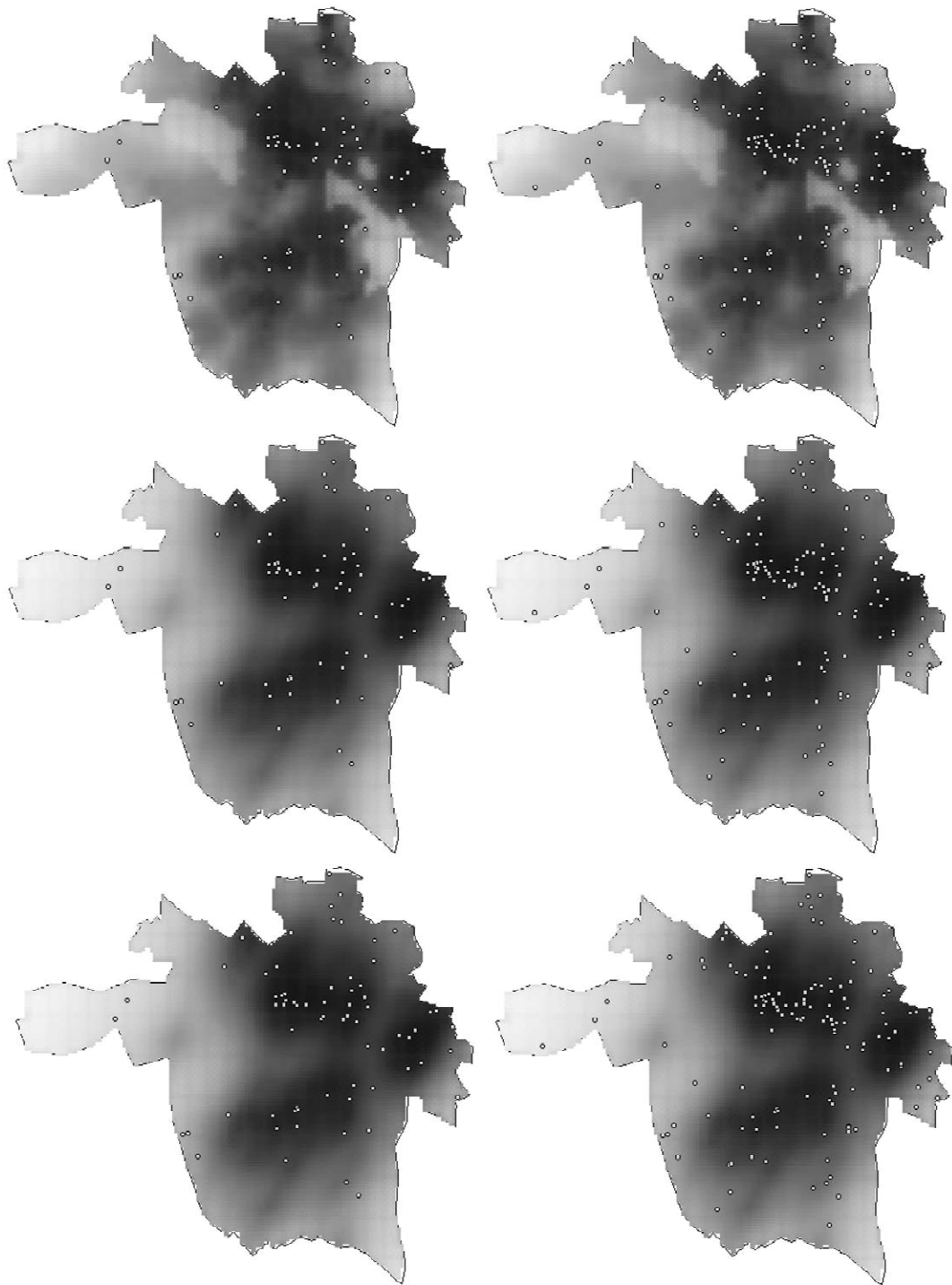


Fig. 3. GMM (upper), WPK (middle), and FPK (lower) versions of the proposed model calibrated on July 7–20 data and tested on July 21–27 data (left) and July 21–August 3 data (right).

proposed model captured more spatial variation than either the WPK or FPK versions. This explains the results in Tables 8–11 where the GMM version is picked as the best model for every weekly or biweekly prediction scenario.

## 5. Conclusion

In this paper, we have described a newly developed space–time prediction model for crime points and evaluated it on breaking and entering burglary point data from Richmond, VA. The proposed model is shown to be more effective than the best of current “hot spot” methods. Some important characteristics of this approach include:

- inclusion of measurable features that are useful for prediction;
- identification of the features with the most predictive or explanatory power; and
- presentation of forecasts through probability density estimates over space and time.

Our prediction modeling can be integrated into an interactive shared information and decision support system such as ReCAP (Brown, 1998) to aid crime fighting in an automated fashion.

## Acknowledgements

This project was partially supported by grant no. NIJ 98-LB-VX008 awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. Points of view in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice.

## Appendix A. Model component estimation

This section addresses the four tasks listed in Section 3 for estimating individual components of our transition density model.

### A.1. Partition event feature data

Intuitively, the number  $C$  of the clusters in the key feature space corresponds to the number of distinct sets of criminal preferences. Unless we have clusters 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 accommodate local covariance structures in the component density models that we will examine 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 size  $n$ , a hierarchical clustering algorithm generates a succession of  $n$  partitions  $P_0, P_1, \dots, P_{n-1}$ , where  $P_0, P_1, \dots, P_{n-1}$  contain  $n, n-1, \dots, 1$  cluster(s), 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 variants of the algorithm. We will not delve into the details and the interested reader is referred to Everitt (1991) for an introduction. The stopping rule that we use is a revision of Mojena (1977). Let  $\alpha_j$  be the shortest distance between any two clusters in the partition  $P_j$  ( $j=0, 1, \dots, n-1$ ). Then revised rule is to stop merging clusters further and select the first partition  $P_j$  satisfying

$$\alpha_{j+1} > \bar{\alpha}_j + k \cdot s_{\alpha_j} \quad (\text{A.1})$$

where  $\bar{\alpha}_j$  and  $s_{\alpha_j}$  are the mean and unbiased standard deviation of  $\alpha_0, \alpha_1, \dots, \alpha_j$ , 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 (1977). The rationale of these rules is to look for significant “jump” in the  $\alpha$  series.

### A.2. Estimate first-order spatial transition density and spatial interaction probabilities

We consider two classes of models for estimating the first-order spatial transition density. Both classes play roles in modeling data from multiple underlying

categories and sources. The first class is called *finite mixture distributions* (e.g., Everitt & Hand, 1981; McLachlan & Basford, 1988; Titterton, Smith, & Makov, 1985). These distributions are superpositions of component distributions. A finite mixture probability density function (or mass function in the case of discrete sample space) has the form

$$f(\mathbf{x}; \boldsymbol{\pi}, \boldsymbol{\Theta}) = \sum_{j=1}^C \pi_j f_j(\mathbf{x}; \boldsymbol{\theta}_j) \tag{A.2}$$

where  $\pi_j > 0, j = 1, 2, \dots, C, \pi_1 + \pi_2 + \dots + \pi_C = 1, \boldsymbol{\pi} = [\pi_1 \dots \pi_C]', \boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_C]. f_j(\mathbf{x}; \boldsymbol{\theta}_j)$  is the  $j$ th component density with the set  $\boldsymbol{\theta}_j$  of parameters and  $\pi_1, \pi_2, \dots, \pi_C$  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 was done in the previous subsection—partitioning event feature data  $\{\mathbf{x}_i; i = 1, 2, \dots, n\}$ .

Two aspects need to be addressed further in order for us to generate a density estimate by Eq. (A.2). First, further assumptions need to be made on the functional form of the component densities  $f_j(\mathbf{x}; \boldsymbol{\theta}_j)$  ( $j = 1, 2, \dots, C$ ). For a continuous feature space (where all features are continuous variables) we use Gaussian mixture models (GMM), where  $f_j(\mathbf{x}; \boldsymbol{\theta}_j), j = 1, 2, \dots, C$ , 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 assume 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 parameters  $\boldsymbol{\pi} = [\pi_1 \dots \pi_C]'$  and  $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 \dots \boldsymbol{\theta}_C]$ . We use a numeric maximum likelihood algorithm known as Expectation-Maximization (EM) algorithm (see, for example, Dempster, Laird, & Rubin, 1977). This algorithm first calculates these parameters with respect to the clusters in the feature space partition, and then updates them iteratively until the log likelihood  $L = \sum_{i=1}^n \log f(\mathbf{x}_i; \boldsymbol{\pi}, \boldsymbol{\Theta})$  converges to a stationary point.

The second class of models we use to estimate the first-order spatial transition density belongs to non-

parametric techniques and was introduced by Marchette, Priebe, Rogers, and Solka (1996). They are collectively called *filtered kernel estimators* (FKE) and take the form

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^C \frac{\rho_j(\mathbf{x}_i)}{|\mathbf{H}_j|} K(\mathbf{H}_j^{-1}(\mathbf{x} - \mathbf{x}_i)) \tag{A.3}$$

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

$$0 \leq \rho_j(\mathbf{x}) \leq 1 \text{ and } \sum_{j=1}^C \rho_j(\mathbf{x}) = 1 \tag{A.4}$$

for all  $\mathbf{x}$ , are *filtering functions*. Local bandwidth matrices contain posterior parameter settings that enforce localized smoothness. The filtering functions are prior weights over variations of local smoothness. We only consider a special case of Eq. (A.3) for our purpose where we set  $\mathbf{H}_j = \text{diag}[h_{j1} \dots h_{jp}]$ ,  $j = 1, 2, \dots, C$ , where  $h_{jl}$  ( $j = 1, 2, \dots, C; l = 1, 2, \dots, p$ ) is a local bandwidth for the  $l$ th dimension  $[\mathbf{x}]_l$  of the  $j$ th locally varied region of support. 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 the standard multivariate Gaussian density function. To generate a density estimate by Eq. (A.3), we need to specify the filtering functions as well as the local bandwidths. Suppose the data  $\{\mathbf{x}_i; i = 1, 2, \dots, n\}$  have been partitioned into  $C$  clusters  $\Omega_1, \Omega_2, \dots, \Omega_C$ . 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

$$\rho_j(\mathbf{x}) = \pi_j g_j(\mathbf{x}) / g(\mathbf{x}), \quad j = 1, 2, \dots, C \tag{A.5}$$

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

$$\rho_j(\mathbf{x}) = \mathbf{1}_{\{\mathbf{x} \in \Omega_j\}}, \quad j = 1, 2, \dots, C \tag{A.6}$$

We term the FPK estimators with the filtering functions defined by Eq. (A.6) *weighted product kernel (WPK) estimators*. Let  $n_j$  be the number of data points in cluster  $\Omega_j$ . The local bandwidths are estimated by using local data in each cluster. To wit,

$$\hat{h}_{jl} = \left( \frac{4}{p+2} \right)^{1/(p+4)} \hat{\sigma}_{jl} n_j^{-1/(p+4)},$$

$$l = 1, 2, \dots, p, \quad j = 1, 2, \dots, C \tag{A.7}$$

where  $\hat{\sigma}_{jl}$  is the standard deviation of the  $l$ th variable  $[\mathbf{x}]_l$  estimated from the unidimensional local data set  $\{[\mathbf{x}_i]_l; \mathbf{x}_i \in \Omega_j\}$ . Notice that these bandwidth estimates are optimal in the Approximate Mean Integrated Square Error (AMISE) sense when they are used to fit Gaussian product kernel estimators to the local data sets  $\{\mathbf{x}_i \in \Omega_j\}$ ,  $j = 1, 2, \dots, C$ , 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 simultaneously specified. Spatial interaction probabilities are estimated using these “local” models. When a finite mixture distribution is involved, spatial interaction probabilities are given as

$$\Pr\{\mathbf{x}_{n+1} \in \mathcal{X}^{(j)} | \mathcal{X}_n^{(j)}\} = \pi_j f_j(\mathbf{x}_{n+1}; \boldsymbol{\theta}_j) / f(\mathbf{x}_{n+1}; \boldsymbol{\pi}, \boldsymbol{\Theta}),$$

$$j = 1, 2, \dots, C \tag{A.8}$$

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

$$\Pr\{\mathbf{x}_{n+1} \in \mathcal{X}^{(j)} | \mathcal{X}_n^{(j)}\} = \hat{f}_j(\mathbf{x}_{n+1}) / \hat{f}(\mathbf{x}_{n+1}),$$

$$j = 1, 2, \dots, C \tag{A.9}$$

where

$$\hat{f}_j(\mathbf{x}_{n+1}) = \frac{1}{n} \sum_{i=1}^n \frac{\rho_j(\mathbf{x}_i)}{|\mathbf{H}_j|} K(\mathbf{H}_j^{-1}(\mathbf{x}_{n+1} - \mathbf{x}_i)),$$

$$j = 1, 2, \dots, C \tag{A.10}$$

### A.3. Estimate second-order spatial transition densities

The third task is to estimate second-order spatial transition densities. The models we choose for these densities maintain continuity in parallel with the ordering of inter-event geographic distances and/or that of inter-event temporal distances. Such orderings reflect additional assumptions on site selection behavior. First, given that two geographic locations have the same set of feature values, it is 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 takes into account the assumption regarding the *lingering period to resume act*. We describe these models below.

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

$$\psi_n^{12}(\mathbf{s} | D_n^{(j)}, T_n^{(j)}, t) = \varphi_m(\mathbf{s} | \mathbf{s}_1, \dots, \mathbf{s}_m)$$

$$= \frac{\lambda^2}{2\pi m} \sum_{i=1}^m e^{-\lambda \|\mathbf{s} - \mathbf{s}_i\|} \tag{A.11}$$

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  $\mathbf{s}_i$  ( $i = 1, 2, \dots$ ,

$m$ ). 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  $t_1, t_2, \dots, t_m$ . Based on this model, we postulate that the second-order spatial transition density for cluster  $j$  takes on the form

$$\begin{aligned} \psi_n^{12}(\mathbf{s}|D_n^{(j)}, T_n^{(j)}, t) &= \eta_m(\mathbf{s}|\mathbf{s}_1, \dots, \mathbf{s}_m, t_1, \dots, t_m, t) \\ &= \frac{\lambda^2}{2\pi \sum_{i=1}^m e^{-\tau(t-t_i)}} \sum_{i=1}^m e^{-\lambda\|\mathbf{s}-\mathbf{s}_i\|-\tau(t-t_i)} \end{aligned} \quad (\text{A.12})$$

For both Eqs. (A.11) and (A.12), we can numerically solve for the maximum likelihood estimates of the parameters (i.e.,  $\lambda$  in Eq. (A.11),  $\lambda$  and  $\tau$  in Eq. (A.12)). The interested reader is referred to Fiksel (1984).

#### A.4. Estimate geographic-space feature density

The fourth and last task is to estimate the geographic-space feature density when appropriate and possible. In general, this requires sampling over the study region. For example, we obtain feature values for sample locations chosen uniformly and independently over the study region. We then fit a density function to these sample values using either finite mixture or filtered kernel method.

## References

- Amir, M. (1971). *Patterns in forcible rape*. Chicago: University of Chicago Press.
- Baldwin, J., & Bottoms, A. (1976). *The urban criminal: A study in Sheffield*. London: Tavistock Publications.
- Block, C. (1995). STAC hot-spot areas: A statistical tool for law enforcement decisions. In Block, C. R., Dabdoub, M., & Fregly, S. (Eds.), *Crime analysis through computer mapping*. Washington, DC: Police Executive Research Forum, p. 20036.
- Brantingham, P., & Brantingham, P. (1975). Spatial patterns of burglary. *Howard Journal of Penology and Crime Prevention*, 14, 11–24.
- Brantingham, P., & Brantingham, P. (1984). *Patterns in crime*. New York: Macmillan Publishing.
- Brown, D. E. (1998). The Regional Crime Analysis Program (ReCAP): A framework for mining data to catch criminals. *Proceedings of 1998 IEEE International Conference on Systems, Man, and Cybernetics*, 2848–2853.
- Capone, D., & Nichols, W. (1976). Urban structure and criminal mobility. *American Behavioral Scientist*, 20, 199–213.
- CensusCD+maps, Version 2.0 (1998). GeoLytics, East Brunswick, NJ.
- Cressie, N. A. C. (1993). *Statistics for spatial data*. New York: Wiley.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B*, 39, 1–38.
- Diggle, P. J. (1983). *The statistical analysis of spatial point patterns*. London: Academic Press.
- Everitt, B. S. (1984). *An introduction to latent variable models*. London: Chapman & Hall.
- Everitt, B. S. (1991). *Cluster analysis*. 3rd ed.. London: Edward Arnold.
- Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman & Hall.
- Fiksel, T. (1984). Simple spatial-temporal models for sequences of geological events. *Elektronische Informationsverarbeitung und Kybernetik*, 20, 480–487.
- Jefferis, E. (1998). A multi-method exploration of crime hot spots. *Presentation at the Annual Meeting of the Academy of Criminal Justice Sciences, Albuquerque, NM, March 10–14, 1998*.
- 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.
- Levine, N. (1998). "Hot Spot" analysis using *CrimeStat* kernel density interpolation. *Presentation at the Annual Meeting of the Academy of Criminal Justice Sciences, Albuquerque, NM, March 10–14, 1998*.
- Marchette, D. J., Priebe, C. E., Rogers, G. W., & Solka, J. L. (1996). Filtered kernel density estimation. *Computational Statistics*, 11, 95–112.
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.
- Milligan, G. W., & 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.
- Molunby, 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*. New York: Macmillan.
- Repetto, T. A. (1974). *Residential crime*. Cambridge, MA: Ballinger.
- Rossmo, D. K. (1993). Target patterns of serial murders: A methodological model. *American Journal of Criminal Justice*, 17(2), 1–21.

- Rossmo, D. K. (1996). Targeting victims: Serial killers and the urban environment. In O'Reilly-Flemming, T. (Ed.), *Serial and mass murder: Theory, research, and policy*. Toronto: Canadian Scholars Press.
- Scarr, H. A. (1973). *Patterns in burglary*. 2nd ed.. Washington, DC: U.S. Department of Justice.
- Scott, D. W. (1992). *Multivariate density estimation*. New York: Wiley.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

**Biographies:** Dr. Hua LIU works as a Senior Software Engineer with CSG Systems, a leading provider of billing and customer care software. Prior to CSG, he was a Member of Technical Staff with Lucent Technologies. His industrial experience includes research and development of state-of-the-art communication software, and large-scale systems modeling, control and optimization.

He has conducted research in the areas of inductive modeling, intelligent systems, probabilistic models, and discrete-event simulation. He holds a BS degree from Tianjin University, China, and MS and PhD degrees from University of Virginia, all in Systems Engineering.

Dr. Donald E. BROWN is Professor and Chair of the Department of Systems and Information Engineering at University of Virginia. He is also Director of the Critical Incident Data Analysis Center and Editor of the IEEE Transactions on Systems, Man, and Cybernetics, Part A. He is a fellow of the IEEE and the 2002 winner of the Norbert Wiener Award from the IEEE Systems, Man, and Cybernetics Society for outstanding contributions to research and education in systems engineering. He is a recipient of the Millennium Medal from the IEEE. He received his PhD from University of Michigan, the MS and ME degrees from the University of California, Berkeley, and the BS degree from the U.S. Military Academy, West Point.