A SPADAC Technical White Paper

28 Sept 2009

Revision 2.1

Geospatial Preference Models in Signature Analyst®

Jason R. Dalton
Michael D. Porter

7921 Jones Branch Dr, Suite 600
McLean, VA 22102
Phone: 703-893-3500
Fax: 703-893-8131
info@spadac.com

Geospatial Preference Models in Signature Analyst®

Jason R. Dalton and Michael D. Porter

28 Sept 2009
Revision 2.1

Abstract

Signature Analyst®is a spatial decision support system that attempts to describe the relationship between Intelligent Site Selection (ISS) event locations and a set of environmental factors using Geospatial Preference (GSP) models. This paper describes the statistical aspects of the GSP modeling approach. Through a flexible additive structure, GSP models can represent complex processes yet maintain interpretability. Furthermore, by attempting to model in the decision space of the actors, the GSP models can provide significant improvement over standard spatial modeling techniques. Evaluation techniques, factor metrics, and event clustering approaches are also covered in more detail.

Contents

1 Introduction to Geospatial Preference Modeling

We describe a methodology for modeling events occurring within a geographic space. These events are being initiated, in a certain geographical region, by actors with complex decision making capabilities. While the actors are making spatial decisions (site selection), they are using more than just spatial coordinates. The decisions are being made based upon a complex combination of the attributes and features of one location compared to the attributes and features of all other available locations, with spatial coordinates providing an index to map a particular site to it's corresponding feature set.

Geospatial preference (GSP) modeling refers to the methodological framework specifically designed to understand the decision making process of the actors as well as predict where future events may occur. GSP models treat the events as coming from an intelligent site selection (ISS) process, a special class of point process arising from spatial decision making. The actors are considered *intelligent* because they make their site selection decisions based upon their preferences or perceived utility of the site. The preferences are based upon some attributes or features associated with the site. The GSP methodology attempts to incorporate the actors' preferences by including the features that are thought to influence (or are associated with the features that influence) the ISS process.

The GSP models are therefore constructed from data (usually GIS layers) related to the attributes and features of the locations in the geographical area of interest (AOI). The objective is to identify the information layers that are responsible for (or associated with) the event site selection, understand their relative importance, and construct likelihood contours for future events.

A reliable forecast method for criminal and terrorist actors would improve resource allocation for patrols, sensors, and counter-measures. Improved resource allocation results in an enhanced ability to detect, classify, and mitigate threats before they endanger people. Furthermore, discovering the tactics, techniques, and procedures (TTP) of the actors would contribute to the understanding of the environmental conditions that are conducive to a successful site selection, helping guide environmental changes where possible.

When an actor selects a site, we assert that it is not the coordinate values of that location, but rather the salient properties of that location, to include environmental factors, infrastructure, demographics, topography, etc. For more details on the use and discovery of these features, see the Signature Analyst®User Guide.

1.1 Intelligent Site Selection and Environmental Theory

We consider an intelligent site selection (ISS) process one in which actors judiciously select the locations and times to initiate events according to their preferences or perceived utility of those locations and times. In other words, the actors have a choice set (or opportunity space) consisting of spatial locations and times and an unknown number of events to initiate. We propose that the preferences and indeed rate of event initiation is guided by some attributes or features associated with the elements in the opportunity space. Furthermore we assume that by capturing the information from the features that are being considered by the actors will lead to "better" models than modeling based on spatial coordinates alone.

This approach for explaining criminal or terrorist site selection behavior is based on environmental criminology theory. These theories seek to describe the motivations and acts of crime based on the general features of one's environment, which can directly relate to the environment of the criminal or crime scene [1, 6, 2]. In other words, the criminals and terrorists choose the location of the crimes/attacks based upon some attributes or features of the location [7]. Additionally, some environmental conditions will encourage or discourage crime, such as poor economic conditions, proximity to police stations, bars and nightclubs [5]. All of these theories point to the explanation that there are some features of the locations of the crimes/attacks which are important to criminal decision making. GSP models use this concept to model the event point process by including a location's attributes or feature values along with the spatial information.

The observed ISS point pattern is essentially the realization of a decision making process, and it reflects the choice behavior of the actors. Thus it is not the decision making process alone that is of interest, but also the result of the decisions, that is, the choice of locations and times of event initiation. We would like to capture the features and attributes, termed *factors*, which are influencing the actors' event initiations. However, since they are specific to the individual actors and generally unknown, we must assume that not all of the relevant factors will be included in our models. In the context of crime and terrorism, for each type of crime or terrorist act, we can expect the set of factors that are used for decision making to differ. For example, a burglar might be interested in locations that offer quick escape or hiding while a suicide bomber might choose locations where there are many people. Indeed there might even be several groups of criminals or terrorists that commit the same crime type but have different motivations and risk levels and thus consider different factor sets in their site selection.

A major complication arises in many ISS processes because the actors' strategies are inaccessible. If theory or expert opinion were available, we could capture the factors that are assumed to influence the decision making process or surrogate factors that could be associated (correlated) with the actual factors being considered. However, this might not be a reliable option due to lack of expertise in specific ISS processes and geographic regions. Therefore, our GSP models attempt to *discover* the relevant factors by mining the previous event data. That is, by observing the past event locations and times, and the values of the factors, we can find patterns in the relationship between the event locations and factors. The factors to include in the GSP models are those that appear to most influence the ISS process.

GSP models are applicable to a wide range of ISS process such as urban warfare, retail building selection, animal movement or habitat selection, archaeological dig planning, natural event locations (e.g. forest fire and earthquakes), corrosion growth, and epidemiology. The last three are not *intelligent* in the sense that a rational decision maker is not performing site selection, but rather nature itself or physical laws are guiding the selection. However, since the event locations and times are still dictated by factor values, and occur under uncertainty, these application areas can be considered as well. Tests for change and surveillance (prospective change detection) methods have been developed for ISS processes in [14, 15].

1.2 Feature Space Preference (FSP) models

Methods dealing with general point processes have been developed around environmental phenomena, which assume a smooth spatial dependence model and the lack of discontinuities [9]. These methods perform well for modeling spatial phenomena that have smooth spatial autocorrelation such as sea surface temperature, forest composition, water table depth, and pollution spreading. However, when these techniques have been applied to the social sciences (and ISS processes in particular), the results are not as promising [12].

To address this problem, Brown, Dalton, and Hoyle [3] outlined a new concept in spatial point process forecasting, feature space preference (FSP) models, designed to directly consider the decision making process of the actors. Their purpose of this modeling effort was to gain insight into the actors' selection preferences by observing sites that were selected previously. By looking at the decision space as a set of environmental variables rather than a set of geographic coordinates, a closer fit to the true selection criteria was derived and more accurate models were created.

The FSP method expands the available data through the use of geographic information systems to derive features that describe the environment surrounding each event of interest in an attempt to build an environmental model of the space in which decisions are made. The FSP method proceeds by using non-parametric density estimation techniques to create a multivariate profile of each event in the high-dimensional factor space. The multiple profiles are then combined to give an assessment score directly related to the actors preferences (in the factor space) which can be projected back onto the geographic space producing a choropleth map of the likelihood values in the AOI.

1.2.1 FSP Formulation

Let s be a spatial coordinate (e.g. longitude, latitude) and define $A = \{s : s \in A\}$ as the geographical Area of Interest (AOI) (e.g. city, country, administrative district) where the ISS process is to be modeled. Furthermore specify $X(s) = [X_1(s), X_2(s), \ldots, X_p(s)]$ as a vector of p measurable factors related to the spatial location s. These p factors will be used to derive our preference models. The n observed event locations constitute a subset of the AOI and are specified by s_i $(i = 1, 2, \ldots, n)$.

One of the primary outputs of FSP is an assessment score. The assessment score, $\{z(s) : s \in A\}$ is an estimate of the relative likelihood that a future event will occur at each location in the AOI. This assessment score can be used to maximize utilization or allocate resources to help prevent or mitigate the effects of future events. Therefore FSP builds a statistical model to express the relationship between the factor values $X(s)$ and the assessment score $z(s)$.

FSP constructs a kernel density estimate [13] of the event probability density function (pdf) over each continuous factor. Taking $x_j(s_i)$ to be the j^{th} factor value for event i, the density can be estimated by

$$\hat{f}_j(u) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(u - x_j(s_i)\right) \tag{1}$$

where the kernel $K_h(\cdot)$ is a symmetric, non-negative function that integrates to 1 and h is the smoothing parameter that controls the smoothness of the resulting density estimate, with larger bandwidths corresponding to a smoother estimate. A commonly used kernel is a mean zero Gaussian density function with standard deviation h

$$K_h\left(x\right) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{x^2}{2h^2}\right) \tag{2}$$

Figure 1 shows a kernel density estimation with Gaussian kernels. The final estimate is the sum of $n = 12$ scaled kernels centered at the event locations.

If the factors are categorical, empirical densities are used. Thus

$$\hat{f}_j(u) = \frac{1}{n} \sum_{i=1}^{n} I(x_j(s_i) = u)$$

with $I()$ as the indicator function, is the density estimate for category u.

The final assessment score is the combination of the p univariate density estimates in the form of an additive model

$$z(s) = \frac{1}{W} \sum_{j=1}^{p} w_j \hat{f}_j(x_j(s)) \tag{3}$$

where $W = \sum_{j=1}^{p} w_j$. The factor weights w_j allow each factor to have a varying contribution toward the final assessment score. Setting all weights to 1 (i.e. $w_j = 1\ \forall j$) essentially treats all factors as equally important to the ISS actors. However, this may not be ideal as certain factors may be of minimal or no importance to the actors' site selection preferences. Setting a factor weight to 0 will completely remove that

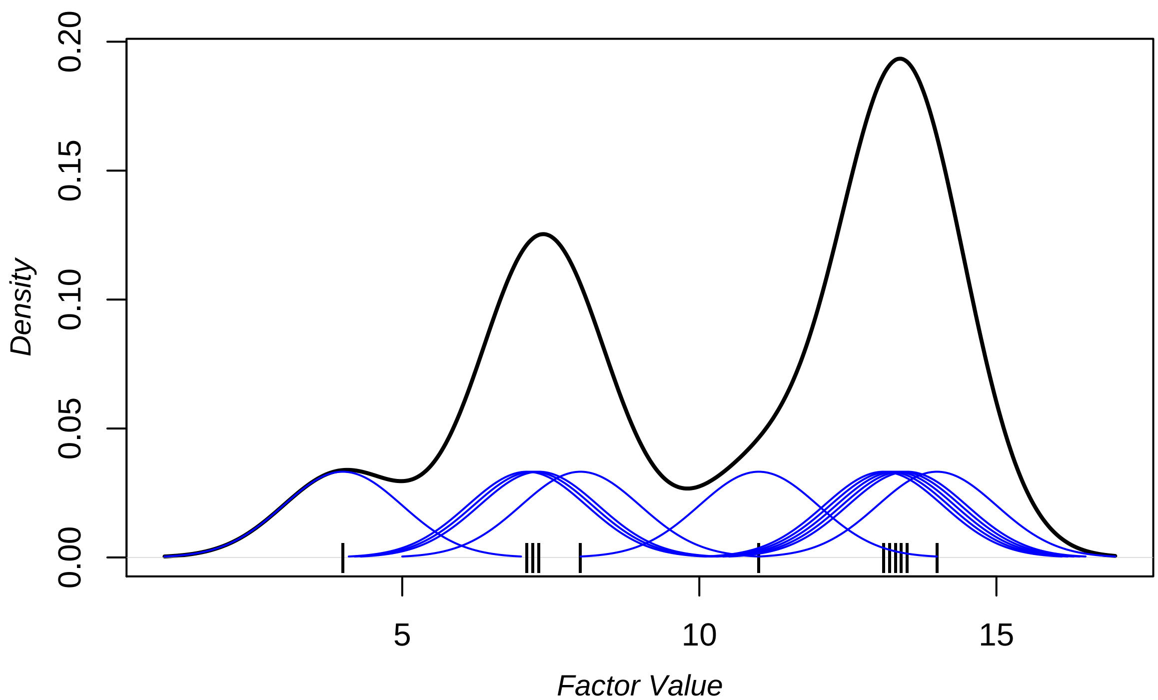

Figure 1: Kernel density estimate. The $n = 12$ event locations are represented by the small tick marks at the bottom of the plot. For each event, a scaled (by 1/12) Gaussian kernel with $sd = 1$ is shown in blue. The resulting density estimate (black curve) is obtained by summing the scaled kernels.

factor from the FSP model (factor selection). Alternatively, setting the factor weights to reflect the contrast between the events and the background environment can help account for the varying importance of each factor.

1.2.2 Contrast Measure

According to the preferences of the ISS actors, some environmental features of a location will be more relevant in their site selection decisions than others. Consequently, we should expect to find that the best models give different weights to the factors. The contrast measure is a statistic that provides insight into the explanatory power of each factor.

Following the assumptions of ISS processes, we presume that the actors made their site selection choices from the available locations defined by the AOI. That is, the selected sites were more preferred than the other possible sites in the AOI. In order to assess the general importance of each factor, we need to examine the distribution of the features values in the AOI.

The distribution of the environment features is estimated by taking a sample of the environment (AOI points) and using kernel density estimation to estimate their pdfs over the available factors. Specifically, assume N locations are taken randomly from the AOI, $s_k : k = 1, 2, \ldots, N$. Then kernel density estimates are obtained from (1) in the same manner as the event pdfs were constructed (except substituting the AOI samples s_k for the event locations s_i). The difference between the event and AOI pdfs is an indication of the importance of a factor. Figure 2 shows the pdfs derived from the events and AOI sample.

The contrast measure is the integral of the absolute difference between the event and AOI pdfs. Formally,

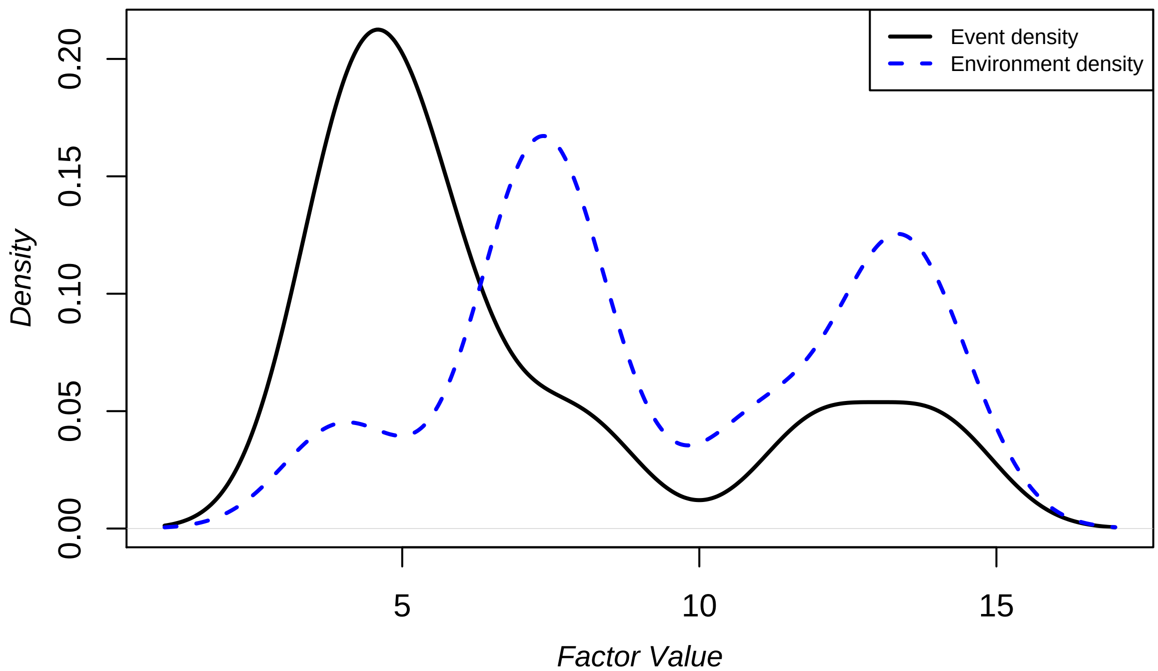

Figure 2: Density estimate of events and random sample from the environment (AOI). Large differences between these two densities signify that the factor may be important to the site selection decisions.

the contrast measure for factor j is

$$c_j = \frac{1}{2} \int |\hat{f}_j^{ev}(x) - \hat{f}_j^{aoi}(x)| dx$$

where $\hat{f}^{ev}$ and $\hat{f}^{aoi}$ refer to the event and AOI estimated pdfs respectively. The contrast measure will take a value between 0 and 1 (i.e. $0 \leq c_j \leq 1$) with larger values corresponding to factors that show a greater discrepancy between event locations and the environment (and are therefore more important). When using the contrast measures as the factor weights ($a_j = c_j$), (3) shows that the factors with a higher degree of contrast are given more influence on the final assessment score.

1.3 Comparison to Spatial Modeling

The types of spatial point processes that the FSP methodology is designed to model is different than the standard spatial processes that more traditional methods address. When these situations prevail, FSP is expected to produce better predictions then existing methods. We will briefly contrast FSP in relationship to spatial density estimates [8] and geographically weighted regression (GWR) [10].

1.3.1 FSP and Spatial Density Estimation

Spatial density estimation (or Hot Spot mapping) assumes that future events will occur in close *geographic* proximity to past events. Consequently they can only be expected to produce good predictions when this condition holds. More specifically, the assumption is that the observed events were generated from a spatial density $f(s)$ that is spatially smooth, has no discontinuities, and that future events will come from the same distribution. If this assumption holds, a spatial kernel density estimate should provide a good forecast of future events. The spatial kernel density estimate for a location $s \in A$ is

$$\hat{f}(s) = \frac{1}{n} \sum_{i=1}^{n} K_h \left(d(s - s_i) \right)$$

where $d(s - s_i)$ is some measure of distance (e.g. euclidean) between locations s and s_i and $K_h(x)$ is a kernel with smoothing parameter of h (see (2) for the Gaussian kernel).

FSP, on the other hand, does not assume that the density over *geographic space* will be smooth, but in fact often expects discontinuities. For example, in urban environments, factors based on physical terrain, infrastructure, demographics, and cultural terrain data can change drastically between locations in close proximity. By assuming a smooth model in the ISS *decision space*, FSP corresponds to smooth models in the preferences of the actors. These variables more closely model an offenders preferences in target selection, and therefore produce a density of preferred target locations, rather than a density of past event locations. This will often lead to better predictions because actors will likely select new locations that have features similar to previous event locations, even if those locations are geographically distant to those previous events.

1.3.2 FSP and Geographically Weighted Regression

FSP assumes the actors' preference structure is stationary over the AOI. This is reasonable when FSP is applied to small or medium AOI's (e.g. operational areas that an analyst might be responsible for). Deriving a model for an entire country or the planet would most likely violate this assumption, as preferences are likely to be somewhat regional and may have small scale variations. This allows the FSP models to use all available historic events in the AOI to build a predictive model. This is important (especially with small numbers of events) as using less data will lead to higher variability in the model outputs.

Alternatively, geographically weighted regression (GWR) assumes that the preferences do change within an AOI. This would imply that actors acting in one region of the AOI have different preferences and utilities than actors who initiate events in another region of the AOI. To account for this, the GWR approach creates a separate model for every location in the AOI. Each model uses weighted data, where the weights correspond to the distance between the model location and the training data. By weighting, GWR essentially does not use all the available training data. This can be beneficial as it can decrease bias if the process is truly non-stationary, but it will also increase the variance of the predictions. This can be especially problematic when there are few training events available, or the events are spaced (geographically) far apart.

GWR produces a varying coefficient [11] GLM model that would give an assessment score

$$z(s) = \beta_0(s) + \sum_{j=1}^{p} \beta_j(s) x_j(s) \tag{4}$$

where the regression coefficients, $\beta(s)$, would be estimated by maximizing a weighted likelihood equation for every location s.

2 Geospatial Predictive Modeling Framework

The FSP model fits into the larger and more general approach to GSP modeling with additive models. Our GSP framework consists of a flexible additive structure that allows a wide variety of models to be developed.

This includes FSP as well as spatial kernel density, geographically weighted regression, naive bayes, logistic regression, gradient boosting, and generalized additive models, to name a few.

Restating the notation, let s be a spatial coordinate (e.g. longitude, latitude) and define $A = \{s : s \in A\}$ as the geographical Area of Interest (AOI) (e.g. city, country, administrative district) where the ISS process is to be modeled. Furthermore specify $X(s) = [X_1(s), X_2(s), \ldots, X_p(s)]$ as a vector of p measurable factors related to the spatial location s. These p factors will be used to derive our preference models. The event locations constitute a subset of the AOI and are specified by s_i $(i = 1, 2, \ldots, n)$.

One of the primary outputs of a GSP model is the assessment score. The assessment score, $\{z(s) : s \in A\}$ is an estimate of the relative likelihood that a future event will occur at each location in the AOI. This assessment score can be used to maximize utilization or allocate resources to help prevent or mitigate future attacks. Therefore GSP builds a statistical model to express the relationship between the factor values $X(s)$ and the assessment score $z(s)$.

Specifically, our GSP framework creates assessment scores of the form

$$z(s) = a_0(s) + \sum_{j=1}^{p} a_j g_j(x_j(s)) \tag{5}$$

for locations s in the AOI. This class of models is additive in the factors, with weights a_j, and some function g of the predictors. The FSP model, given in (3), uses $a_0(s) = 0$, $a_j = w_j/W$, and $g_j = f_j$ with f_j the kernel density estimation given in (1).

2.1 Binary Classification

ISS processes are the result of a site selection decision; the actors specifically choose one location to initiate an event over the others. The responses of such a process are of two kinds, locations where events occurred and locations where they did not occur. Therefore, to facilitate our modeling of an ISS process, we will structure our formulation as a binary classification problem. That is, we will select N_0 locations in A to represent the environment (or opportunity space) which we term *negatives* and use the N_1 event locations as the *positives*. This is equivalent to case-control data in epidemiology.

The data used for modeling can be represented as

$$D = [(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N), (X_{N+1}, Y_{N+1}), \ldots, (X_{N+n}, Y_{N+n})]$$

where

$$Y_i = Y(X_i) = \begin{cases} 1 & \text{if the } i^{th} \text{ sample is a positive (event)} \\ 0 & \text{if the } i^{th} \text{ sample is a negative (environment)} \end{cases}$$
$$X_i = [X_{i,1}, X_{i,2}, \ldots, X_{i,p}] \text{ is a vector of } p \text{ factor variables}$$

The goal is to express the relationship between the binary response variable and predictor variables by a mathematical formula. Specifically, we want to estimate the posterior probability $\Pr(Y = 1|X = x)$ for all x values in the AOI. This is an estimate of the probability that a point at location s with predictor values $X(s) = x$ is an event.

Bayes Theorem can provide a more descriptive view of what we are trying to solve. To simplify notation, represent $p(x) = \Pr(Y = 1|X = x)$ as the posterior probability, $f_k(x) = \Pr(X = x|Y = k)$ as the class conditional density or likelihood, and $\pi = \Pr(Y = 1)$ as the prior probability. Bayes Theorem expresses the relationship between the posterior, likelihood, and prior

$$p(x) = \frac{\pi f_1(x)}{\pi f_1(x) + (1-\pi) f_0(x)} \tag{6}$$

Since $p(x)$ is a probability it must be between $[0, 1]$. This restriction limits many direct modeling approaches. A common remedy is to construct models in a transformed space, by using transformations such as the logit. Instead of estimating $p(x) = \Pr(Y = 1|X = x)$ directly, we can transform to estimate the logit,

$$\gamma(x) = \log \frac{p(x)}{1 - p(x)}$$

or log posterior odds. Since $\gamma(x)$ is in the range of $(-\infty, \infty)$, we are no longer bothered by constrained estimators.

Filling in from (6), we get

$$\gamma(x) = \log \frac{\pi}{1-\pi} + \log \frac{f_1(x)}{f_0(x)} \tag{7}$$

We will show that this is in the same form as the assessment scores in (5) demonstrating the close relationship between our GSP formulation and Bayesian decision theory. By applying the inverse logit transformation,

$$p(x) = \frac{e^{\gamma(x)}}{1 + e^{\gamma(x)}} \tag{8}$$

the original probability estimates can be recovered.

2.2 Other GSP Models

Besides the FSP model, there are many other types of models that fit into out GSP framework. We outline a few in this section that use the binary classification formulation.

2.2.1 Naive Bayes (NB)

The Naive Bayes (NB) model gets its name from making the naive assumption (because it is rarely true) that the factors are independent. This assumption would imply that the multivariate class conditional density can be obtained by multiplying the marginal densities

$$\begin{aligned} f_k(x) &= \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p | Y = k) \\ &= \prod_{j=1}^{p} \Pr(X_j = x_j | Y = k) \\ &= \prod_{j=1}^{p} f_{kj}(x_j) \end{aligned}$$

(using $f_{kj}(x_j) = \Pr(X_j = x_j | Y = k)$) making estimation much easier.

There are various methods to estimate the marginal class conditional densities. We choose to use non-parametric kernel density estimation, when appropriate, giving

$$\hat{f}_{kj}(u) = \frac{1}{N_k} \sum_{i=1}^{N_k} K_h \left(u - x_j(s_i)\right)$$

with $k \in \{0, 1\}$ where N_k is the number of events of type k. For the events (i.e. $k = 1$), this is exactly the FSP estimate (1). For $k = 0$, this gives as estimate of the density of the factor values in the AOI.

This leads to an estimate the logit

$$\hat{\gamma}(x) = \log \frac{N_1}{N_0} + \sum_{j=1}^{p} \log \frac{\hat{f}_{1j}(x_j)}{\hat{f}_{0j}(x_j)} \tag{9}$$

To put this in the form of the GSP framework (5), use $a_0(s) = \log\left(N_1/N_0\right)$, $a_j = 1$, and $g_j(x_j(s)) = \log\left(\hat{f}_{1j}(x_j(s))/\hat{f}_{0j}(x_j(s))\right)$.

2.2.2 Biased Sampling Algorithm (BSA)

The motivation behind the biased sampling algorithm (BSA) is that the site selections can be viewed as biased random samples from the AOI. The sampling is biased because some locations are more attractive or preferred by the actors that others. The bias function for location s

$$\delta(s) = \frac{f_1(x(s))}{f_0(x(s))}$$

is the likelihood ratio of event density w.r.t the environment density at that location.

Following NB, BSA estimates the marginal densities separately, but instead of making the independence assumption, it finds an optimal weighting to combine the estimates. This gives the BSA assessment scores

$$z(s) = \beta_0 + \sum_{j=1}^{p} \beta_j \, \log\left(\hat{\delta}_j(x_j(s))\right) \tag{10}$$

where $\hat{\delta}_j(u) = \hat{f}_{1j}(u)/\hat{f}_{0j}(u)$ and the coefficients β are fit with maximum likelihood. This corresponds to a logistic regression model with transformed predictors. BSA can be advantageous to NB because it attempts to find an optimal combination of the factors, whereas NB blindly treats all factors equally.

2.2.3 Logistic Regression (LR)

Logistic Regression (LR) is a type of generalized linear model (GLM) that can also be represented by the GSP framework. LR assumes that a location's attractiveness can be modeled as a linear combination of the factor values. This gives an assessment score

$$z(s) = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_j(s) \tag{11}$$

where the regression coefficients β are estimated by maximum likelihood with the iteratively reweighed least squares (IRLS) algorithm.

2.2.4 GAM and Boosting

Generalized Additive Models (GAM) [11] and componentwise Gradient Boosting [4] attempt to directly model the component functions $g_j(x_j(s))$ using nonparametric smoothing. Overfitting is controlled by constraining the smoothness of the functions. The general form of these models is

$$z(s) = \hat{a}_0(s) + \sum_{j=1}^{p} \hat{R}_j(x_j(s)) \tag{12}$$

where $\hat{R}_j(u)$ is some smooth function across factor j.

2.3 Summary of Models

Our GSP framework is a very general modeling structure that supports a wide variety of models. Table 1 summarizes some of the models that fit into the GSP framework. Besides the FSP and Spatial Kernel models, the other models produce an assessment score on $(-\infty, \infty)$. These can be transformed to $[0, 1]$ through the inverse logit transformation (8). When the objective is to identify the top $(\alpha \times 100)\%$ search region, this will have no effect since this is a monotonic transformation. Likewise, $a_0(s)$ will also have no effect on the top search region as long as $a_0(s) = a_0$ is a constant over all $s \in A$ (this will not be the case for Spatial Kernel and GWR).

Method	Eq.	$\alpha_0(s)$	α_j	$g_j(x_j(s))$
FSP	(3)	0	w_j/W	$\hat{f}_{1j}(x_j(s))$
Naive Bayes	(9)	$\log(N_1/N_0)$	1	$\log\left(\hat{f}_{1j}(x_j(s))/\hat{f}_{0j}(x_j(s))\right)$
BSA	(10)	$\hat{\beta}_0$	$\hat{\beta}_j$	$\log\left(\hat{f}_{1j}(x_j(s))/\hat{f}_{0j}(x_j(s))\right)$
Logistic Regression	(11)	$\hat{\beta}_0$	$\hat{\beta}_j$	$x_j(s)$
GAM	(12)	$\log(N_1/N_0)$	1	$\hat{R}_j(x_j)$
Spatial Kernel	(1)	$\hat{f}(s)$	0	0
GWR	(4)	$\hat{\beta}_0(s)$	$\hat{\beta}_j(s)$	$x_j(s)$

Table 1: Summary of GSP models of the form $z(s) = a_0(s) + \sum_{j=1}^{p} a_j g_j(x_j(s))$

Our GSP models can also be viewed as as an ensemble method. Each *signature*, g_j can be considered a submodel constructed from only one of the factors. The final GSP model is a weighted sum of these signatures, where the weights, a_j indicate how much influence each submodel has on the final output. The additive structure can still handle complicated functional forms by augmenting the predictors. That is, additional factors can be developed from special functions of the original factors. For example, it may be helpful to consider spatial coordinates, quadratic terms, and interaction terms as additional factors.

3 Model Output and Evaluation

GSP models for ISS processes are used for predicting where future events will occur and understanding the actors preferences and site selection behavior. The model building can be an iterative process that includes choosing the specific model for an assessment, performing factor selection to eliminate redundant or uninformative factors, and determining the events to use for model training. A model evaluation technique will help guide this process and allow an analyst to find the best model for a certain ISS process. As the objectives of the GSP modeling process are often to disrupt or influence the site selections, the performance evaluation of spatial site selection models requires some modification to the standard approaches. In addition, we describe some factor metrics and clustering techniques to gain understanding of the ISS decision making process.

3.1 Prediction Evaluation

One goal of the GSP models is to provide a prediction or forecast of where future events will occur. For criminal and terrorist processes, this would allow patrols, allocations, searches, surveillance, interventions, or other responses to be focused on regions in the AOI where the most activity is expected to occur. Due to the inherent cost and availability constraints, resources and manpower cannot be allocated to the entire AOI. Therefore, a good prediction will identify the regions that will contain the most future events, as well as be able to accurately predict the number or fraction of events that will occur in those regions.

This is similar to the goals of earthquake prediction, where a decision needs to be made concerning when and where to issue an alarm or warning. We specifically focus on the situation where resources can be allocated to only $(\alpha \times 100)\%$ of the AOI at any one time period (this region is called the α *search region*). In this situation, the assessment scores are used to identify the α search region and evaluation is made on how many events actually occur in this region.

Let $B(\alpha) \subseteq A$ be the α search region. This is the upper level set

$$B(\alpha) = \{u \in A : z(u) \geq z_\alpha\}$$

where z_α is the smallest threshold such that $B(\alpha)$ is at most $(\alpha \times 100)\%$ of the AOI.

The identified $B(\alpha)$ provides an estimate of the region where the most future activity will occur, but it does not provide any information about how much activity is to occur there. If n_α events are observed in the search region $B(\alpha)$ (out of a total of N events), the proportion of events occurring the search region are denoted by $y(\alpha) = n_\alpha/N$ and is referred to as the *accuracy* of the search region.

3.1.1 Area Reduction Curves (ARC)

Finding the search regions $\{B(\alpha)\}$ and accuracies $\{y(\alpha)\}$ for all $0 \leq \alpha \leq 1$ gives the Area Reduction Curve (ARC), which graphically displays the amount of AOI that needs to be searched to capture a certain proportion of events shown in Figure 3. The ARC, which is similar to a rotated receiver operating characteristic (ROC) curve, gives an estimate of the performance of the assessment for any sized search region.

One measure of the overall performance of a model is the area under the ARC (similar to the AUROC for ROC curves)

$$Y = \int_0^1 y(a)\, da$$

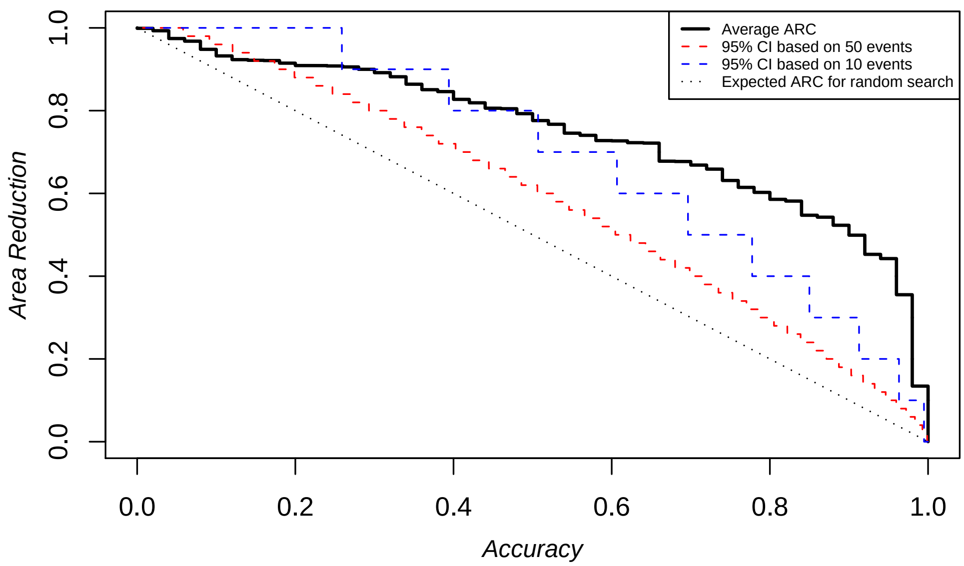

Figure 3: Area Reduction Curve (ARC). This shows the accuracy, or proportion of events captured, plotted against the area reduction, or $1 - \alpha$ search region. For a completely random search (the worst case), you would expect the ARC to be the diagonal given by the dotted line. The dashed lines show 95% one-sided confidence intervals (CI) based on a random search. The red line is based on 50 events and is much tighter the blue line based on only 10 events.

This estimates the probability that the assessment score at a randomly selected event is greater than the assessment score for randomly selected location in the AOI. Y will be in the range of $[0, 1]$ with a 1 denoting a perfect assessment and 0.5 the expected performance of a random search.

3.1.2 Cross-Validation

To prevent an inflated view of performance, evaluation is usually carried out on a test set, or data that was not used to fit the model. Cross-validation is an approach that uses multiple test sets to reduce variability in the performance estimate. For K-fold cross-validation, the events are partitioned into K equally sized subsets. For every fold, one subset is used as the testing set and the remaining subsets are used for training. This produces K different training and test sets.

Suppose there are N total events available for modeling. For fold k, there will be $N_0(k)$ training events and $N_1(k)$ test events. Based on the training events, a model produces an assessment $\{z_k(s) : s \in A\}$ and corresponding search regions $\{B_k(\alpha) : 0 \leq \alpha \leq 1\}$. The hold out test set is used to construct the ARC curve $y_k(\alpha)$ and the number of test events, $n_k(\alpha)$, in the α search region.

The average accuracy is obtained from averaging the K cross-validation accuracies

$$y(\alpha) = \frac{1}{N} \sum_{k=1}^{K} n_k(\alpha) = \frac{n(\alpha)}{N} \tag{13}$$

where $n(\alpha) = \sum_{k=1}^{K} n_k(\alpha)$ is the total number of events that occurred in the set of α search regions. Taking the area under this average ARC produces a summary statistic that can be used to evaluate model performance and compare competing models.

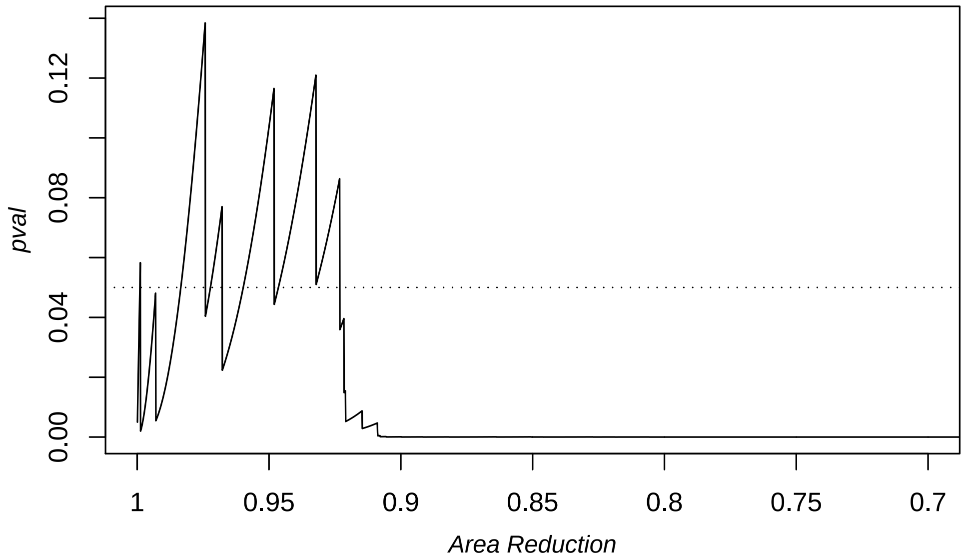

Figure 4: Instead of viewing the ARC with the confidence intervals, a $pval$ curve can be viewed instead. The $pval$ curve is a plot of the $pval$ against area reduction. Low values of $pval$ correspond to significant area reduction levels. In this example, the $pval$'s drop to very low values when the area reduction is below 0.92. This shows that the GSP models lead to significantly (from a statistical perspective) better search regions than a random search strategy when the search regions are greater that $1 - 0.92$ or 8% of the AOI.

3.1.3 Accuracy Significance

Even the average ARCs can be somewhat misleading because they don't reveal the variability of the estimate. The accuracy score, $y(\alpha)$ is the expected proportion of future events that will occur in search region $B(\alpha)$. This is actually a binomial random variable and is subject to sampling variation. Under cross-validation, the uncertainty of this value is a function of the total number of training events as well. An ARC curve constructed from 100 events will be much more reliable that one constructed from 20.

One way to determine the quality of the ARCs is to compare it's performance to a random search. A random search implies that $B(\alpha)$ is constructed by taking random subsets of the AOI and corresponds to the worst possible performance. Under a random search strategy, the number of events in the search region will follow the binomial distribution

$$n(\alpha) \sim Bino(N, \alpha)$$

with expected value of $N\alpha$.

This distributional information allows the creation of confidence intervals (CI) for the expected ARC under a random search strategy. The GSP models produces significantly better search regions than a random strategy for area reduction values in which the average ARC exceeds the CIs. Figure 3 shows one-sided 95% confidence intervals based on 10 and 50 events. The average ARC in this figure exceeds the CI based on 50 events for Area Reduction values less than 0.9 (i.e. α search region values greater than 0.1). If, however, the average ARC curve was build from only 10 events, the CI is much wider and the ARC doesn't exceed it until Area Reduction values less than around 0.8 (i.e. α search region values greater than 0.2). This shows the difficulty of attributing a significant performance gain to GSP models constructed from few training events.

Instead of the confidence intervals, p-values can be used to assess the significance of a model's accuracy improvement. After observing $n_{obs}(\alpha)$ events in the set of α search regions, the p-value is

$$pval = \Pr(N(\alpha) \geq n_{obs}(\alpha)|\mathcal{H}_0) = 1 - \Pr(N(\alpha) \leq n_{obs}(\alpha) - 1|\mathcal{H}_0)$$

where $n_{obs}(\alpha) - 1$ is used because the counts are discrete. These values can be obtained from any statistical software or textbook tables. For a small enough p-value (e.g. $pval \leq 0.05$) the $\mathcal{H}_0$ (the null hypothesis of random search) is rejected and the evidence points to the given model being better than a random search (at search proportion α). Figure 4 is a plot of the $pval$ for various area reduction values. This shows more clearly that the model will produce significantly better search regions than the random search for area reductions less than 0.92 or when the more than 8% of the AOI can be searched.

3.1.4 Cost Curves

The ARC provides an estimate of the probability that future events will occur in certain regions of the AOI. This, in turn indicates the potential impact that any actions or mitigation efforts might have on event initiation. In order to justify resource expenses or determine the optimal size of the search region the ARC curves can be converted to cost curves.

A cost curve replaces the accuracy with a cost measure (alternatively one could consider a loss, utility, or benefit measure instead of direct cost). The cost will likely be a function of several aspects. First, the cost can be a function of the amount of effort or resources needed for the size of the search area (i.e. α). The cost can also be a function of the number of events missed (i.e. $\mathbb{E}[N_{new}(1 - y(\alpha))]$) and/or the number of events captured (i.e. $\mathbb{E}[N_{new}y(\alpha)]$) . This would allow an expected cost function to be derived from the performance estimates in the ARCs and expected number of new events.

3.2 Factor Metrics

Besides the assessment score (and corresponding search regions) the influence of the factors on event initiation is also an important part of GSP modeling. While identifying causal relationships are usually beyond reach, *factor metrics* can be used to gain an understanding of the association between event locations and factor values. A factor metric is an indication of how much influence a factor has on the assessment score.

The factor metrics refer to two aspects of the GSP models (5), the weights a_j and the signatures g_j. The weights give an indication of the amount of information (for prediction) which is contained in a factor. This is a global measure and assesses the overall usefulness of a factor for predicting over the entire AOI. The signatures, alternatively, provide the local information about the relationship between event initiation and factor values. The signatures will be larger for locations that have factor values close to the training event factor values (i.e. hot spots in the decision space). Figure 5 shows the signatures for a model with four different types of factors. The assessment score for a location is formed from taking the weighted sum of the signature values in the decision space.

It might also be informative to investigate the average signature values for certain subregions in the AOI (such as the event points or α search region). For a region $B \in A$, the average signature is $\int_B g_j(u)\, du$. This provides an indication of how the factors relate to the predictions for these specially chosen regions.

3.3 Event Clustering

The standard GSP models assume that all the training events (and consequently future events) are all from the same distribution (i.e. from the same ISS process). One ISS process can be comprised of numerous actors; the actors of a common ISS process are all the individuals or groups that can initiate the same type of

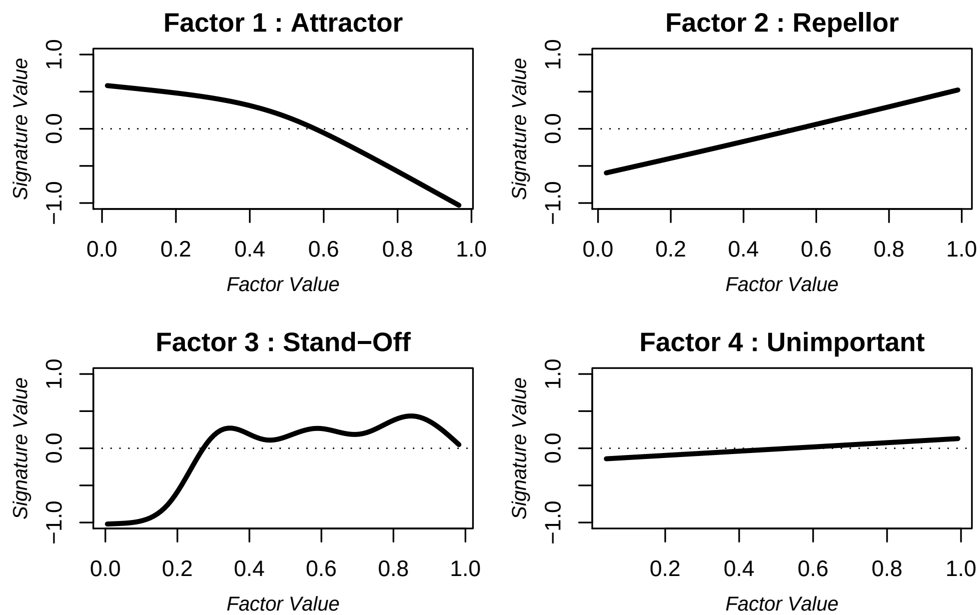

Figure 5: Component plots of the signatures. This shows the signatures for a model with four types of factors (attractor, repellor, stand-off, and unimportant). These plots show how the assessment score is effected by a change in the factor values.

event. If some of these actors have different preferences, or consider different factors, then the resulting GSP models may lose some precision as they attempt to create one model to explain the actions of heterogeneous actors. Better assessments may be available if separate models were developed for different actors.

Dendrograms can be used to identify the heterogeneous groups of actors, even if no actor information is available. Dendrograms are a graphical tree diagram that represents the degree of similarity between the events and clusters of events (see Figure 6). They are the result of a hierarchical clustering algorithm. Agglomerative hierarchical clustering is a sequential procedure that begins by calculating a distance between all events (all events are initially considered to be in their own cluster). The distance will be based on the event's multivariate factor values (e.g. Mahalanobis distance). The events with the smallest distance are then merged to form a new cluster. The distances between the new cluster and the other events is then calculated. There are several ways this distance is calculated e.g. single linkage, complete linkage, or average linkage. This process is repeated until all events are in the same cluster.

The height at which the clusters are merged is proportional to the distance between the two sets of events in the clusters. So if there are homogeneous clusters of events in the data the dendrogram will show the tightness of the clusters and the distance between the clusters. These are also useful for identifying outlier events.

4 Summary

Geospatial Preference (GSP) models are a flexible family of models for understanding Intelligent Site Selection (ISS) processes. Their additive structure make interpretation easy, but through the unconstrained specification of the fitting functions have the ability to model very complex processes. By using a large set of environmental predictors, GSP models attempt to discover patterns in the decision space of the actors

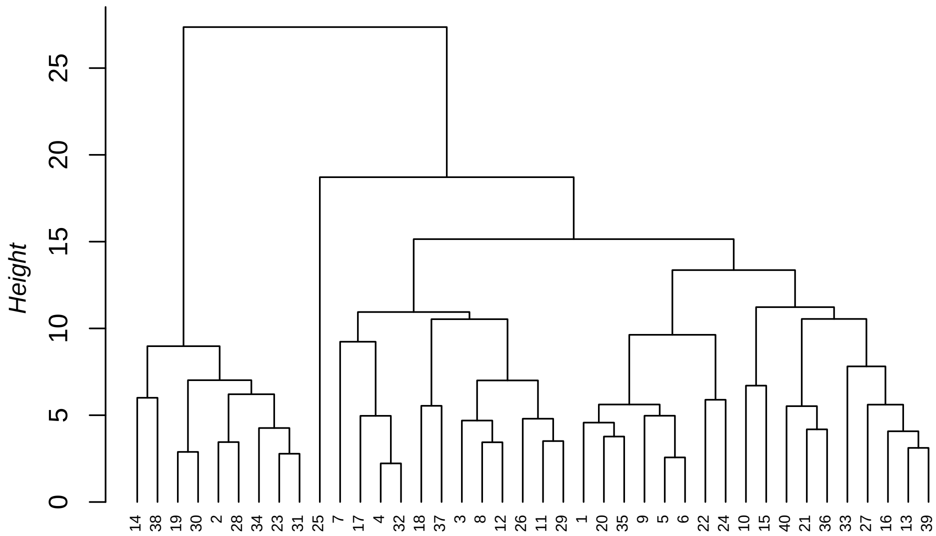

Figure 6: Dendrogram from 40 events. There appears to be three or four main clusters. One cluster is an individual event ($\#25$).

yielding improvements over standard spatial modeling techniques.

The improvement from GSP models can be evaluated through Area Reduction Curves (ARC) with easy to calculate confidence intervals and significance tests. By describing how the signatures add up to the assessment score, factor metrics can aid in the understanding of the relationship between the factors and event locations. Dendrograms and other event clustering techniques can identify if the data includes any clusters or outliers, helping to reform and improve the models.

References

[1] L. Anselin, J. Cohen, D. Cook, W. Gorr, and G. Tita. *Spatial analyses of crime*, volume Volume 4: Measurement and Analysis of Crime and Justice of *Criminal Justice 2000*, pages 213–258. National Institute of Justice, Washington, DC, 1 edition, 2000.

[2] P. J. Brantingham and P. Brantingham. *Environmental Criminology*. Sage Publications, Beverly Hills, CA, 1981.

[3] D. Brown, J. Dalton, and H. Hoyle. Spatial forecast methods for terrorist events in urban environments. *Intelligence and Security Informatics*, 3073:426–435, 2004.

[4] P. Bühlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

[5] R. J. Bursik. Social disorganization and theories of crime and delinquency: Problems and prospects. *Criminology*, 26(6):519–551, 1988.

[6] J. Byrne and R. Sampson. *The Social Ecology of Crime*. Springer-Verlag, New York, 1986.

[7] L. E. Cornish and R. Clarke. *The Reasoning Criminal: Rational Choice Perspectives on Offending*. Springer-Verlag, New York, 1986.

[8] P. Diggle. *Statistical Analysis of Point Patterns*. 2nd edition, 2003.

[9] Dueker and Kjerne. *Environmental modeling with GIS*. 1989.

[10] A. Fotheringham, M. Charlton, and C. Brunsdon. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning*, 30(11):1905–1927, 1998.

[11] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[12] N. Lavigne. *Mapping Crime: Principles and Practice*. 1999.

[13] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:10651076, 1962.

[14] M. D. Porter. *Detecting space-time anomalies in point process models of intelligent site selection*. PhD dissertation, University of Virginia, Department of Systems and Information Engineering, 2006.

[15] M. D. Porter and D. E. Brown. Detecting local regions of change in high-dimensional criminal or terrorist point processes. *Computational Statistics and Data Analysis*, 51(5):2753–2768, 2007.