ISeB (2007) 5:255–274
DOI 10.1007/s10257-007-0045-1

ORIGINAL ARTICLE

Discrete choice analysis of spatial attack sites

Michael A. Smith · Donald E. Brown

Published online: 20 March 2007

Abstract This paper presents an algorithm for the complete specification of multinomial discrete choice models to predict the spatial preferences of attackers. The formulation employed is a modification of models previously applied in transportation flow and crime analysis. A breaking and entering crime data set is employed to compare the efficacy of this model with traditional hot spot models. Discrete choice models are shown to perform as well as, or better than such models and offer more interpretable results.

Keywords Spatial choice · Multinomial choice · Feature selection · Preference specification · Target selection

1 Introduction

In the contemporary, increasingly mobile society, a large number of the choices individuals face are spatial choices. The broad topic of this paper is inferring the preferences that govern specific spatial choice decisions from the details of past decisions.

Familiar instances of spatial choice are informative in setting a context. The choice of a fueling station along one's daily commuting route is a spatial choice problem in which the decision maker may consider location relative to

M. A. Smith · D. E. Brown (✉)
Department of Systems and Information Engineering,
University of Virginia,
P.O. Box 400747, Charlottesville, VA 22904, USA
e-mail: brown@virginia.edu

M. A. Smith
e-mail: mas3f@alumni.virginia.edu

traffic patterns, price differences, perceived crime level, and countless other characteristics of candidate solutions. Deciding among restaurants for dinner before seeing a theater production is a spatial choice problem. In this decision, one might consider proximity to the theater and available parking, in addition to aspatial attributes such as cuisine and price. Making a residential real estate purchase is a highly spatial problem in which one considers neighborhood characteristics such as population demographics, school districts and likely resale value. All of these problems differ in the extent to which spatial attributes influence the decision making process, but they share several common properties.

Formalizing the shared properties (Fotheringham and O'Kelly 1989) creates a groundwork for discussing spatial choice problems and a basis for this paper. First, the decision produces a binary response; each alternative is either selected or not selected. Second, each alternative is described by a vector of attributes. These attributes may include spatial and non-spatial components. Further, they may be relevant to the decision maker and the decision making process. Finally, the relative spatial placement of each alternative influences its attractiveness. It is this final property that distinguishes spatial from aspatial choice problems.

1.1 Rationale

There are many spatial choice problems that are not encountered by the typical person in daily life, but have a dramatic impact on life. The locations criminals choose to operate in impact the lives of those living in the same areas. The coordinates of hidden improvised explosive devices in combat zones is the product of a spatial choice problem, as is the site at which a suicide bomber chooses to detonate. Clearly, these decisions impact the victims of the attacks. These are all examples of spatial choice problems where mining the spatial preferences of the decision maker and predicting decision outcomes can enable defensive or counteractive measures, thereby improving public safety. The pursuit of this predictive ability is the motivation for this work.

1.2 Objectives

The objective of this paper is to define an algorithm for the *objective* application of multinomial discrete choice models to spatial point patterns which are driven by human decision makers. The facility inherent in the algorithm will be demonstrated on an urban crime data set. This application shows that such models can outperform traditional models that are not based on discrete choice theory, and that discrete choice models are more interpretable by the analyst.

2 Conceptual review

This section discusses the statistical methods necessary to apply the model specification algorithm proposed in Sect. 3.

2.1 Multinomial choice modeling

To formalize the discrete choice models employed in this paper, some notation is required[1]. Let C be the universal choice set, the set of all alternatives to a specific decision. By convention, $|C| = J$. This section discusses models where $J > 2$. Binary models are simpler, but not relevant to many spatial choice applications.

The utility of each element of C is modeled as a random variable, the sum of a systematic part and a random disturbance.

$$U_i = V_i + \varepsilon_i \tag{1}$$

i is used as an index into the choice set.

Random utility theory prescribes that the probability of selection for an individual alternative is

$$\begin{aligned} P(i) &= Pr\{U_i \geq U_j, \forall j \neq i \in C\} \\ &= Pr\{V_i + \varepsilon_i \geq V_j + \varepsilon_j, \forall j \neq i \in C\} \\ &= Pr\{\varepsilon_j \leq V_i - V_j + \varepsilon_i, \forall j \neq i \in C\} \end{aligned} \tag{2}$$

The most commonly applied multinomial choice model is the multinomial logit. This model assumes that the disturbances are independent and identically distributed according to the type 1 extreme value distribution. The selection of this distribution is made for analytical convenience and is defended as a approximation of a Gaussian distribution. The next section discusses this distribution and Sect. 2.1.2 describes the multinomial logit model.

2.1.1 Type 1 Extreme value distribution

In the 1920s, the limiting distributions of maximum values in random samples were first described as extreme value distributions (Johnson and Kotz 1970). The type 1 form has the cumulative distribution function

$$F_X(x; \xi, \theta) = \exp\left\{-e^{-\frac{x-\xi}{\theta}}\right\} \tag{3}$$

and density function

$$f_X(x; \xi, \theta) = \theta^{-1} e^{-\frac{x-\xi}{\theta}} \exp\left\{-e^{-\frac{x-\xi}{\theta}}\right\} \tag{4}$$

where ξ is a location parameter and θ is a scale parameter subject to the constraint $\theta > 0$.

[1] The presentation of multinomial discrete choice models here is adapted from (Ben-Akiva and Lerman 1985) and is kept notationally consistent with that work when possible. A notable difference is that the presentation here assumes a single decision maker.

For convenience, for the remainder of this section, the notation $EV_1(\xi,\theta)$ will indicate the type 1 extreme value distribution with location parameter ξ and scale parameter θ.

Three properties of the distribution make it attractive for use in the multinomial choice model. First, it is preserved over linear transformations. That is,

$$\forall a, b \in \mathcal{R}\left\{X \sim EV_1(\xi,\theta) \rightarrow aX + b \sim EV_1\left(a\xi + b, \frac{\theta}{a}\right)\right\} \tag{5}$$

Second, the difference of two type 1 extreme variates is logistically distributed.

$$\left\{X \sim EV_1(\xi_X,\theta), Y \sim EV_1(\xi_Y,\theta) \rightarrow X - Y \sim F_{X-Y}(x-y) = \frac{1}{1+e^{\theta(\xi_X-\xi_Y-(x-y))}}\right\} \tag{6}$$

Third, the distribution of the maximum of a set of type 1 variates is also type 1 distributed.

$$\forall k \in \mathbb{Z}\left\{\left\{X_i : i \in \mathbb{Z}_1^k\right\} \ni X_i \sim EV_1(\xi_i,\theta) \rightarrow \max_{i\in\mathbb{Z}_1^k} X_i 1(\theta^{-1}\ln\left(\sum_{i\in\mathbb{Z}_1^k} e^{\xi_i\theta}\right), \theta)\right\} \tag{7}$$

2.1.2 Multinomial logit

Derivation of the multinomial logit model follows directly from the properties of the type 1 extreme value distribution. The derivation presented is a variant of one in Ben-Akiva and Lerman (1985).

Equation 2 can be rewritten as

$$P(i) = Pr\left\{V_i + \varepsilon_i \geq \max_{j\neq i\in C} V_j + \varepsilon_j\right\} \tag{8}$$

By assuming $\{\varepsilon_i\} \sim EV_1(0,1)$ and exploiting Eq. 5,

$$\forall j \in C\left\{V_j + \varepsilon_j \sim EV_1(V_j, 1)\right\} \tag{9}$$

and, using Eq. 7

$$\max_{j\neq i\in C} V_j + \varepsilon_j \sim EV_1\left(\ln\sum_{j\neq i\in C} e^{V_j}, 1\right) \tag{10}$$

Letting j^* be the index of the maximum, we can separate the systematic and stochastic components such that

$$\begin{aligned} \varepsilon_{j^*} &\sim EV_1(0,1) \\ V_{j^*} &= \ln \sum_{j \neq i \in C} e^{V_j} \end{aligned} \tag{11}$$

Returning to Eq. 8,

$$\begin{aligned} P(i) &= Pr\left\{V_i + \varepsilon_i \geq V_{j^*} + \varepsilon_{j^*}\right\} \\ &= Pr\left\{\left(V_{j^*} + \varepsilon_{j^*}\right) - (V_i + \varepsilon_i) \leq 0\right\} \end{aligned} \tag{12}$$

Noting Eq. 6, the selection probability becomes

$$\begin{aligned} P(i) &= \frac{1}{1 + e^{V_{j*} - V_i}} \\ &= \frac{e^{V_i}}{e^{V_i} + e^{V_{j*}}} \\ &= \frac{e^{V_i}}{e^{V_i} + \exp\left\{\ln \sum_{j \neq i \in C} e^{V_j}\right\}} \\ &= \frac{e^{V_i}}{\sum_{j \in C} e^{V_j}} \end{aligned} \tag{13}$$

It is notable that the ratio of selection probabilities for any two elements of the choice set is defined completely by those alternatives' systematic utilities.

$$\begin{aligned} \frac{P(i)}{P(j)} &= \frac{e^{V_i} / \sum_k e^{V_k}}{e^{V_j} / \sum_{k \in C} e^{V_k}} \\ &= \frac{e^{V_i}}{e^{V_j}} \\ &= e^{V_i - V_j} \end{aligned} \tag{14}$$

This property, typically referred to as the independence of irrelevant alternatives (IIA) property, is a burdensome constraint that is often violated in choice problems. Several approaches have been taken to remove the property, one targeted at spatial choice problems is presented in the next section.

The multinomial logit derivation presented did not impose a functional form on the systematic utility component. The typical form is linear in parameters, which offers tractability in computation and interpretation. The remainder of this paper assumes a linear in parameters form.

$$V_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} = \boldsymbol{\beta}' \mathbf{x}_i \tag{15}$$

The derivation presented did assume a location parameter, $\xi = 0$, and scale parameter, $\theta = 1$, for the distribution of the stochastic utility component. These values were chosen to reduce the notational complexity of the presentation. Alternate values for these parameters change the scale of the utility, but do not change the selection probabilities that the model produces. As a result, in applications of the multinomial logit model, arbitrary selection of these parameters is a necessity. The values selected here are typical, for reasons of convenience.

2.1.3 Hierarchical multinomial choice

Fotheringham (1986) observed that for spatial choice problems, the formulation presented in Eq. 13 and the IIA property it maintains imply that all elements of the choice set are evaluated equally. In such problems, if the decision maker restricts the choice set by only evaluating particular clusters of alternatives, this assumption is violated.

Following the presentation in Fotheringham et al. (2000), let $M \subseteq C$ be the restricted choice set in which all alternatives are evaluated. Equation 2 is modified to include the likelihood that an alternative is in this restricted set.

$$P(i) = Pr\{U_i \geq U_j + \ln Pr\{j \in M\}, \quad \forall j \neq i \in C\} Pr\{i \in M\} \tag{16}$$

Following the derivation presented in the previous section and maintaining the assumption that $\{\varepsilon_i\} \sim EV_1(0,1)$, the selection probability becomes

$$P(i) = \frac{Pr\{i \in M\} e^{V_i}}{\sum_{j \in C} Pr\{j \in M\} e^{V_j}} \tag{17}$$

Intuitively, this model is a modification of the multinomial logit in which the utility of each alternative is weighted by the likelihood that it was evaluated. It is referred to in this paper as the *spatial hierarchy* model.

The models presented in this paper use the spatial hierarchy model, with evaluation likelihood determined via Parzen kernel density estimation.

3 Model specification

As stated in Sect. 1.2, the objective of this paper is to define an algorithm for the objective application of multinomial discrete choice models to spatial point patterns. It is important to emphasize that the rationale is influenced by the type of patterns under consideration. The algorithm presented in this section assumes that the random variable under consideration represents a spatial attack. In such applications, the goal of the modeler is to predict the locations of future attacks. This specification further assumes that the predictions will be used to spatially deploy limited defensive resources. Therefore, the efficiency of the model is measured by how well it identifies the highest risk subset of the geographic space.

This section details the algorithm used to fully specify a spatially based multinomial discrete choice model. It discusses choice set generation, feature selection, the use of a metric to measure evaluation hierarchies, and model estimation. The algorithm is exercised in Sect. 4.

It is not the goal of this algorithm to predict when or if an attack will occur. The models it constructs estimate the probability that an attack will occur in an arbitrary subset of the geographic space, given that it will occur somewhere within the space.

It is assumed that this algorithm is applied in a supervised learning context, to the training set after an observation data set has been partitioned into training and test subsets. In discrete choice problems, this is a partition of the vector of selections. In the context of spatial attack, this means that a subset of attacks are held out as a test set. The model assumes temporal stationarity, therefore the partition into test and training set need not occur based on an epoch.

In describing the algorithm's steps, elements of notation are used to remove ambiguity. As before, let C be the choice set, containing spatial locations indexed as $\{i : i \in \mathbb{Z}_1^J\}$. Let the candidate predictor set $\Phi = \{\phi_i : i \in \mathbb{Z}_1^L\}$. Each element of the choice set is then described by a vector of attributes $\mathbf{x}_i = [x_{i,\phi_1}, \ldots, x_{i,\phi_L}]$. Let $\gamma_{\phi_j} = [x_{1,\phi_j}, \ldots, x_{J,\phi_j}]$. The data set also contains a vector of selections[2], $\mathbf{s} = [s_1, \ldots, s_N]$ where $s_i \in C$. Let $S = \{\alpha : \alpha \in \mathbf{s}\}$.

3.1 Choice set generation

The multinomial choice model requires a discrete choice set, but the modeler often faces a problem where the decision maker is choosing a single point in a continuous two dimensional space. To discretize the continuous geographic space, a regular grid is laid over the area of interest. The grid cell centroids constitute the choice set. Attacks in the dataset are mapped to the geographically nearest element in the choice set.

The resolution of the grid used must be fine enough to capture the variance of spatial features. Determining adequate resolution depends on the nature of features used. If features are aggregated areally, such as demographic data reported by census tract, then a grid resolution finer than the aggregation area is unnecessary. Similarly, if features are proximity measures, such as distances to key locations, then the inter-centroid spacing should be smaller than the minimum distance between these locations.

3.2 Feature selection

The feature selection component of the algorithm synthesizes the results of agglomerative clustering and the use of Fisher's discriminant ratio to measure feature separability. The result is an ordering of candidate predictor variables that attempts to maximize classification performance by minimizing feature correlation.

[2] Note that this formalization permits an element of the choice set to be selected in multiple observations, i.e., it is possible that $s_i = s_j$ for some $i \neq j$.

3.2.1 *Candidate predictor scoring*

Fisher's discriminant ratio, f_r, is a metric used to score each feature's predictive ability. The choice set is partitioned into elements that have been selected, S, and those that have not, $C - S$. For any set A, let

$$\hat{\mu}_{\phi_j,A} = \frac{1}{|A|}\sum_{i\in A} x_{i,\phi_j} \tag{18}$$

be an estimate of within set mean. Similarly, define within set variance as

$$\hat{\sigma}^2_{\phi_j,A} = \frac{1}{|A|-1}\sum_{i\in A}(x_{i,\phi_j} - \hat{\mu}_{\phi_j,A})^2 \tag{19}$$

Then

$$f_r(\phi_j) = \frac{(\hat{\mu}_{\phi_j,S} - \hat{\mu}_{\phi_j,\{C-S\}})^2}{\hat{\sigma}^2_{\phi_j,S} + \hat{\sigma}^2_{\phi_j,\{C-S\}}} \tag{20}$$

is a measure of the separation between those elements of the choice set that were and were not selected, with respect to candidate predictor φ_j. The greater the value of f_r for a particular feature, the greater the linear separability of the predictor. By calculating f_r for each candidate predictor an ordering of separability is attained.

3.2.2 *Candidate predictor clustering*

Agglomerative clustering is applied to $\left\{\gamma_j : j \in \mathbb{Z}_1^L\right\}$, to identify groups of features with large intra-group correlation. The goal of such grouping is to avoid selecting multiple features from within a correlated group. This is desirable because a high degree of multicollinearity within the model makes parameter estimates less stable and produces a less interpretable model.

In this application, the similarity metric used is correlation magnitude.

$$s(\phi_i, \phi_j) = \left|\frac{\frac{1}{J}\sum_k (x_{k,\phi_i} - \hat{\mu}_{\phi_i,C})(x_{k,\phi_j} - \hat{\mu}_{\phi_j,C})}{\sigma_{\phi_i,C}\sigma_{\phi_j,C}}\right| \tag{21}$$

Complete linkage agglomeration is used, i.e., in the recursive clustering process, given a set of clusters, the two clusters, A and B, are joined if they collectively minimize the metric

$$s_{CL}(A,B) = \max_{a\in A, b\in B} s(a,b) \tag{22}$$

The output of the clustering procedure is a single cluster, ϕ_1, with defined partitions to create any number of clusters $\{1 \ldots L\}$.

With the cluster structure defined, a specific number of predictors, k, can be selected by partitioning the single cluster into k subclusters or correlation groups, [$\phi_{k,1}$ ... $\phi_{k,k}$]. The features within each correlation group are ordered based on discriminant ratio score, f_r. The candidate predictor with the highest score is used in the model, yielding a model with k predictors, each from different correlation groups.

$$\forall i \in \mathbb{Z}_1^k \left(\phi^*_{\varphi_{k,i}} = argmax_{\phi \in \varphi_{k,i}} f_r(\phi)\right) \tag{23}$$

Let $\mathbf{x}^*$ denote the vector of selected attributes.

$$\mathbf{x}_i^* = \left[x_{i,\phi^*_{\varphi_{k,1}}}, \ldots, x_{i,\phi^*_{\varphi_{k,k}}}\right] \tag{24}$$

The appropriate value of k is discussed in Sect. 3.6.

3.3 Spatial hierarchy

Section 2.1.3 details an approach to weighting the utility of choice set alternatives by the likelihood those alternatives are evaluated. This section presents an optional application of that approach.

As previously, let $M \subseteq C$ be the restricted choice set in which all alternatives are evaluated. Beginning with the assumption that the decision maker evaluates the decision in the geographic space before the feature space, it follows that the estimate that a specific alternative, j, was evaluated, $Pr\,\{j \in M\}$, rely solely on geographic location. To meet this end, a spatial kernel density estimate, $\hat{\lambda}(\cdot)$, is produced using the location of selections, $s_i \in \mathbf{s}$, and parameterized with the the cross validation maximum likelihood bandwidth.

More formally, partition the selection vector, $\mathbf{s}$ into M equally sized vectors, $\left\{\mathbf{t}_j : j \in \mathbb{Z}_1^M\right\}$. Then $Pr\{j \in M\} = \hat{\lambda}(j; h^*)$, where

$$\hat{\lambda}(x; h^*) = \frac{1}{Nh^*} \sum_{s_j \in \mathbf{s}} g\left(\frac{||x - s_j||}{h^*}\right) \tag{25}$$

and $g(\cdot)$ is the standard Gaussian probability density function. By defining $\hat{\lambda}(x; h)_{-i}$ as the cross validation kernel density estimate

$$\hat{\lambda}_{-i}(x; h) = \frac{M}{(N-1)h} \sum_{s_j \notin \mathbf{t}_i} g\left(\frac{||x - s_j||}{h}\right) \tag{26}$$

h^* is chosen such that

$$h^* = argmax_{h \in \mathcal{R}^+} \prod_{i=1}^{M} \prod_{s_j \in \mathbf{t}_i} \hat{\lambda}_{-i}(s_j; h) \tag{27}$$

3.4 Estimation

With k predictors selected, the linear-in-parameters multinomial choice model takes the form

$$P(i) = \frac{e^{\beta_1 x^*_{i,\phi^*_{\varphi_{k,1}}} + \cdots + \beta_k x^*_{i,\phi^*_{\varphi_{k,k}}}}}{\sum_{j \in C} e^{\beta_1 x^*_{i,\phi^*_{\varphi_{k,1}}} + \cdots + \beta_k x^*_{i,\phi^*_{\varphi_{k,k}}}}} \tag{28}$$

Parameters are estimated via maximum likelihood, where the likelihood function is given by

$$\mathcal{L}(\boldsymbol{\beta}; C, \mathbf{s}) = \prod_{s_i \in \mathbf{s}} \prod_{j \in C} P(j)^{\mathbb{1}\{s_i = j\}} \tag{29}$$

with corresponding log likelihood

$$l(\boldsymbol{\beta}; C^{\text{c}}\,\mathbf{s}) = \sum_{s_i \in \mathbf{s}} \sum_{j \in C} \mathbb{1}\left\{s_i = j\right\} \left(\boldsymbol{\beta}' \mathbf{x}^*_j - \ln \sum_{k \in C} e^{\boldsymbol{\beta}' \mathbf{x}^*_k} \right) \tag{30}$$

$l(\boldsymbol{\beta}; C, \mathbf{s})$ has been shown (McFadden 1974) to be globally concave if the data, $\left\{\mathbf{x}^*_i\right\}$, are not multicollinear. Thus, obtaining and solving the system of equations produced by the first derivatives will yield maximum likelihood estimates. As a maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$ is asymptotically normal. The associated Hessian matrix can be used to obtain the variance of the parameter estimates, thus feature coefficients can be tested for significance with a Student's t test.

If the optional estimate of $Pr\{i \in M\}$ presented in Sect. 3.3 is included in the model, estimation is modified by appending $\ln \hat{\lambda}(i)$ to $\mathbf{x}^*{}_i$.

For estimation performed in Sect. 4 of this article, numerical optimization was performed using a variable metric algorithm implemented in the R software package (Team 2004).

3.5 Evaluation

As previously stated, this algorithm is designed under the assumption that the predictions produced will be used to spatially deploy limited resources. To effectively meet this use, the model selected must accurately identify the highest risk subset of the geographic space. With this in mind, the evaluation metric of interest is mean selection site percentile score. Given a set of predictions, $\{P(i): i \in C\}$, the percentile score of an alternative, j, is the proportion of alternatives with estimated selection probability less than or equal to that of alternative j.

$$\pi(j) = \frac{1}{J} \sum_{i \in C} \mathbb{1}\{P(j) \geq P(i)\} \tag{31}$$

Informally, assuming a resource allocation strategy which covers areas with highest selection probability first, 1–$\pi(j)$ is the minimum fraction of the total area that must be covered by resources to insure coverage at the attack site.

To calculate the mean selection site percentile score for a dataset, the percentile score is averaged over all selected alternatives.

$$E[\pi(j)|j \in S] = \frac{1}{N} \sum_{i=1}^{N} \pi(s_i) \tag{32}$$

3.6 Model complexity

The mean selection site percentile score provides an evaluation metric and a basis for making decisions on appropriate model complexity. In Sect. 3.2 it was assumed that the number of desired predictors, k, was known. With an evaluation metric established, k is chosen to maximize the mean selection site percentile score. Following recommendations in Hastie et al. (2001), a series of models is fit, over a range of k. For each model, M-fold cross validation is used to estimate the distribution of selection site percentile score. The chosen model is the most parsimonious model within one standard error of the maximum.

3.7 Algorithm summary

This section presented an algorithm for specifying and estimating a multinomial spatial choice model. That algorithm is summarized by the following steps:

1. Define the set of alternatives, C, by discretizing the geographic space with a regular grid.
2. Define the set of candidate predictor variables, Φ.
3. For each candidate predictor, $\varphi_i \in \Phi$, calculate the separability score, $f_r(\varphi_i)$.
4. Build a cluster of candidate predictors, ϕ_1, using correlation magnitude, $s(\cdot)$ as a similarity metric and complete linkage agglomeration.
5. For a range of values of k

(a) Partition the cluster into k subclusters, $\phi_{k,1}$ ... $\phi_{k,k}$
(b) Select the candidate predictor from each cluster with the maximum separability score $\phi^*_{\varphi_{k,i}}$ (as defined in Eq. 23).

(c) Use M-fold cross validation to estimate the mean selection site percentile score and the standard error of this statistic for a model which uses the k chosen predictors, $\left[\phi^*_{\varphi_{k,1}}, \ldots, \phi^*_{\varphi_{k,k}}\right]$.

6. Select as k the size of the most parsimonious model within one standard error of the maximum mean selection site percentile score.

4 Application

4.1 Breaking and entering crimes

The dataset of breaking and entering crime analyzed here was obtained from the Richmond City Records Management Center and augmented with publicly available GIS layers to introduce proximity measurements and census information. For each datum, a geographic coordinate and an attribute vector is included. The same data was used for analysis described (Lia and Brown 2003).

The complete dataset consists of all reported breaking and entering crime in the City of Richmond, VA for calendar year 1997. To make the assumption of consistent preferences reasonable, only *residential* breaking and entering crimes occurring in the third quarter of the year were used. This subset contained 637 distinct spatial selections. It was randomly partitioned into test and training subsets in a 1:2 ratio, producing a training set of 425 crime events and a test set of 212.

4.1.1 Choice set definition

The choice set was defined using the following steps. The smallest rectangle that would wholly circumscribe the boundaries of the City of Richmond was the extents of a grid with 100 rows and 100 columns. A spatial intersection operation was then performed to remove those grid cells falling wholly outside the city limits. The remaining 4,895 centroids constitute the choice set. The north–south inter-centroid spacing is 170 m, the east–west inter-centroid spacing is 191 m.

4.1.2 Candidate predictors

The candidate predictor set, Φ, consists of proximity measurements to each of 30 features in the geographic space and 38 demographic profile variables aggregated at the census block group level. It is notable that some geographic features are subsets of other features, e.g., φ_{26} is the minimum distance to any road and φ_{27} is the minimum distance to an interstate. Clearly, $\forall i\{i \rightarrow x_{i,\phi_{26}} \leq x_{i,\phi_{27}}\}$. Similarly, many of the demographic variates are related, e.g., φ_{40} through φ_{45} are the fraction of the population within a specific age group. It follows then that $\forall i,j\left\{i \in C, j \in \mathbb{Z}_{40}^{45} \rightarrow x_{i,\phi_j} \in [0,1] \wedge \sum_{k=40}^{45} x_{i,\phi_k} = 1\right\}$. The presence of such features creates a possibility of extreme multicollinearity if care is not taken in feature selection.

The large number of candidate predictors prevents description of each in this article.

Before any analysis was performed, the predictors were all standardized to zero mean and unit variance.

4.1.3 Feature selection

Using the 425 site training set, separability scores, f_r, were calculated for each of the candidate predictors. Although the selection vector, **s**, contains 425 elements, because many incidents were closest to the same choice set element, $|S| = 319$. Thus, the f_r values indicate the separability between two classes, C–S, with cardinality 4,576, and S, with cardinality 319. The set of candidate predictors, Φ, was clustered using correlation as a similarity metric and complete linkage as the agglomeration method. Figure 1 includes a dendrogram of candidate predictors showing the results of the agglomerative clustering and a barplot showing the range of separability scores. Labels on the figure indicate the order in which predictors are added to the model.

To select the number of predictors used, tenfold cross validation was performed on models fitting a range of 1–23 total predictors. The maximum mean selected percentile score was obtained using twenty predictors, but the two predictor model was within one standard error, and thereby selected. Figure 2 illustrates the estimated performance as a function of model size.

4.1.4 Model estimation and interpretation

A model was refit using the full training set and only the selected predictors, φ_{30} and φ_{66}. A summary of the estimated model is in Table 1.[3]

The model summary shows that both predictors are extremely significant, with p-values less than 10^{-9}.

The coefficient on φ_{30}, a proximity measure, is negative, indicating the feature it represents is attractive. φ_{66} is a density measurement, and the coefficient indicates positive association with selection likelihood. The coefficient magnitudes indicate a similar impact on alternative utility.

[3] The selected predictors, φ_{30} and φ_{66}, represent distance to the nearest federal highway and population density, respectively. The details of the relationship of these predictors to residential breaking and entering crime is outside the scope of this paper and left to criminologists to interpret.

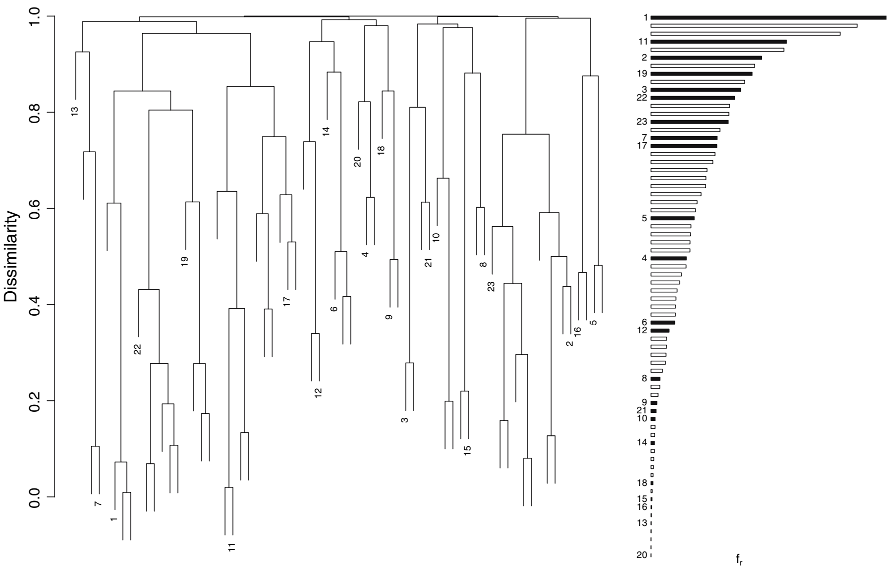


Fig. 1 Breaking and entering: feature selection

Table 1 Breaking and entering: multinomial logit estimation results

Variable	Coefficient estimate	Asymptotic standard error	*t* statistic	*p* value
φ_{30}	–0.563	0.088	–6.411	3.87×10^{-10}
φ_{66}	0.505	0.026	19.474	$< 2 \times 10^{-16}$
AIC = 6,788				
BIC = 6,796				

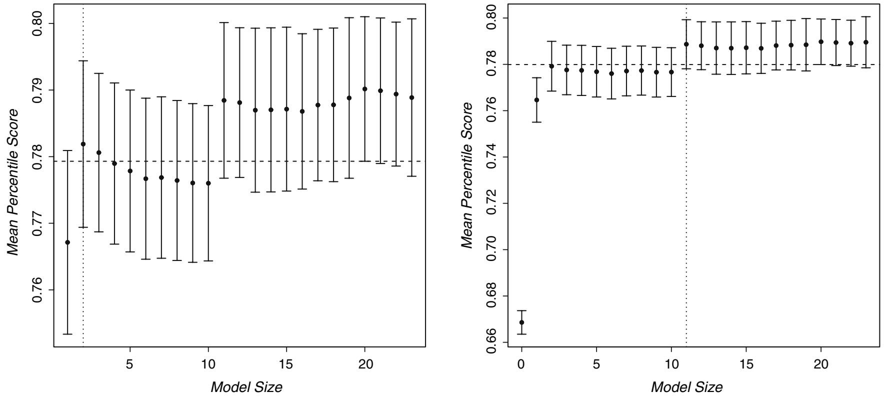
4.1.5 Alternative model: spatial hierarchy

An additional model was estimated using the spatial hierarchy specification described in Sect. 3.3. The spatial kernel density estimate, $\hat{\lambda}$, was created with the maximum likelihood bandwidth of 4.80 km.

The feature selection process (illustrated in Fig. 2) resulted in 11 predictors being used, a much larger set than the standard multinomial logit model. The additional predictors introduced are

$$[\phi_{13}, \phi_{18}, \phi_{19}, \phi_{38}, \phi_{46}, \phi_{52}, \phi_{53}, \phi_{58}, \phi_{62}]$$

The model summary is in Table 2. It indicates that the spatial hierarchy term is significant. φ_{30} exhibits significance in the standard multinomial logit model, but does not in the larger spatial hierarchy model. The *p* value moved from a highly significant value, less than 10^{-9}, to the insignificant 0.189. This change suggests multicollinearity with other features in the model.



(a) Multinomial Logit **(b)** Spatial Hierarchy

Fig. 2 Breaking and entering: complexity versus performance

4.1.6 Model comparison

A geographic hot spot model[4] was estimated for comparison purposes. This kernel density model is equivalent to the likelihood weighting term in the spatial hierarchy model, as presented in Sect. 3.3. As estimated, it used an isotropic Gaussian kernel with maximum likelihood bandwidth of 4.80 km.

Figures 3, 4, and 5 are choropleths of the percentile score over the geographic space for the hot spot, standard multinomial choice, and hierarchical

Table 2 Breaking and entering: spatial hierarchy estimation results

Variable	Coefficient estimate	Asymptotic standard error	*t* Statistic	*p* value
$\log(\hat{\lambda})$	0.595	0.245	2.430	0.016
φ_{13}	–0.056	0.055	–1.008	0.314
φ_{18}	–0.627	0.099	–6.308	7.3×10^{-10}
φ_{19}	–0.017	0.056	–0.296	0.767
φ_{30}	–0.140	0.106	–1.317	0.189
φ_{38}	0.029	0.060	0.478	0.633
φ_{46}	–0.154	0.072	–2.146	0.033
φ_{52}	0.391	0.079	4.975	9.6×10^{-7}
φ_{53}	0.172	0.070	2.468	0.014
φ_{58}	–0.073	0.055	–1.327	0.185
φ_{62}	–0.024	0.036	–0.651	0.516
φ_{66}	0.445	0.041	10.911	$< 2 \times 10^{-16}$
AIC = 6,711				
BIC = 6,760				

[4] Readers interested in the specification and application of hot spot models using kernel density estimation are referred to Fotheringham et al. (2000). The use of kernel density estimation to model crime is extended to incorporate non-spatial attributes in Lia and Brown (2003).

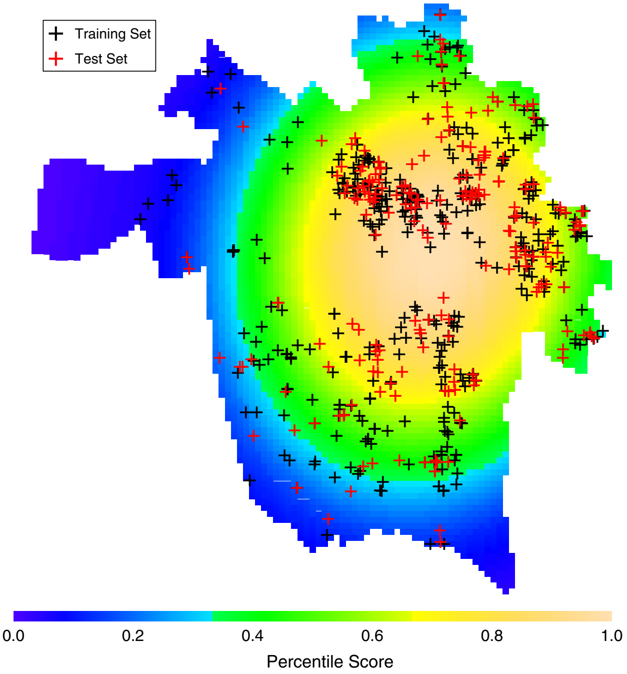


Fig. 3 Breaking and entering: hot spot model

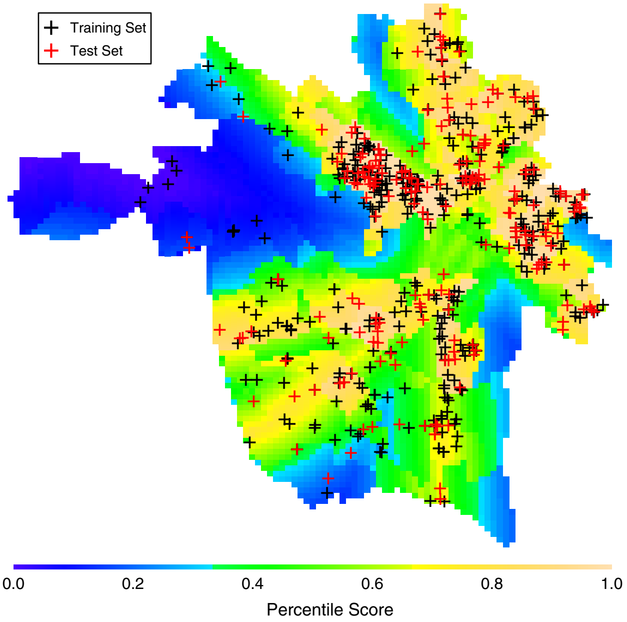


Fig. 4 Breaking and entering: multinomial logit model

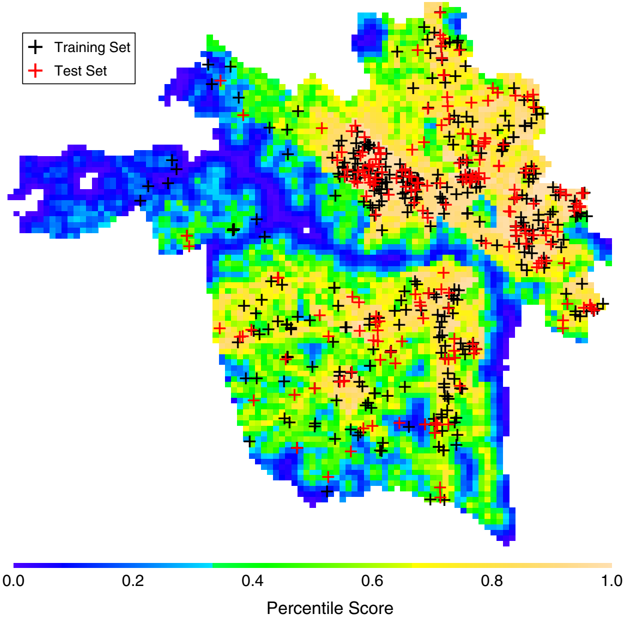


Fig. 5 Breaking and entering: spatial hierarchy model

multinomial choice models respectively. These figures also include points indicating the location of event sites in the training and test sets.

Informally, the most striking difference among the choropleths is that the hot spot model produces much smoother contours than the discrete choice models. This difference reflects the large kernel density bandwidth estimated for the hot spot model and the restriction of that model to the geographic space. For this application, the use of feature space attributes causes the selection probabilities, within some local geographic regions, to vary much more dramatically in the discrete choice models than they do in the hot spot model.

To illustrate the change in distribution of test set predictions, Fig. 6 shows boxplots of the percentile scores by model. The boxplots suggest that the

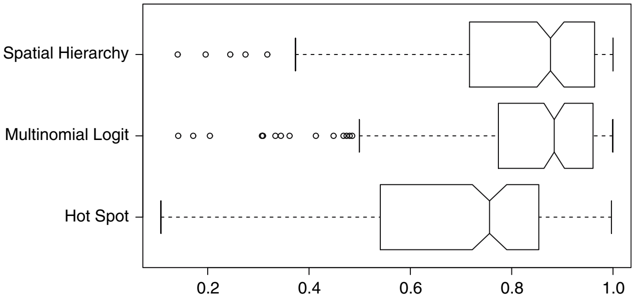


Fig. 6 Breaking and entering: percentile score distribution by model

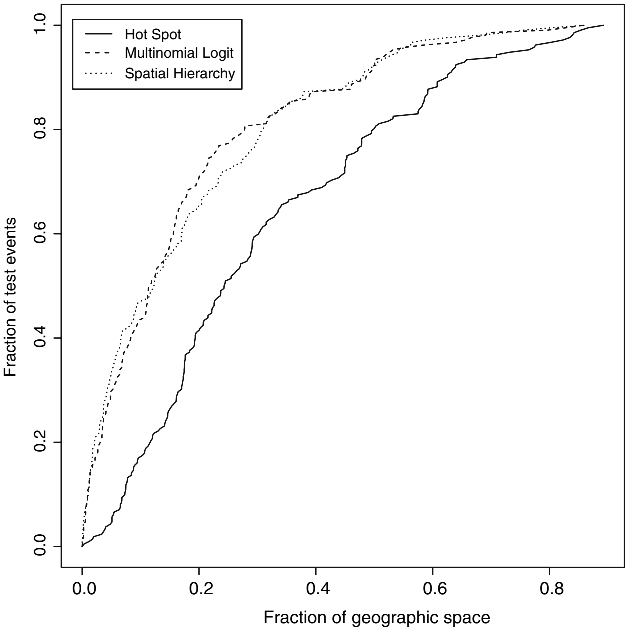


Fig. 7 Breaking and entering: area versus accuracy trade-off

multinomial logit model offers a reduction in variance and an improved median selection site percentile score relative to the hot spot model. The spatial hierarchy model also improves on the hot spot model, though to a lesser extent.

Figure 7 shows the trade-off between the fraction of the geographic space at or above a percentile score and the fraction of test events at or above that percentile score. It shows that the discrete choice models dominate the hot spot model over the complete trade-off range. Further, the difference between the discrete choice models is small throughout, suggesting that the additional complexity and sacrifice of interpretability in the spatial hierarchy model is undesirable.

The large test set makes it possible to apply a paired *t* test on the percentile score distributions to evaluate their relative effectiveness. The null hypothesis is that the mean difference in selection site percentile score is equal to zero. The alternate hypothesis is that the mean difference is greater than zero. Table 3 summarizes the results of these tests, using π_{KDE}, π_{MNL}, and π_{SH} to indicate the hot spot, standard multinomial log, and spatial hierarchy per-

Table 3 Breaking and entering: model comparison results

Test	Test statistic	*p* value
$\pi_{MNL} - \pi_{KDE}$	–9.71	$< 2.2 \times 10^{-16}$
$\pi_{SH} - \pi_{KDE}$	–10.74	$< 2.2 \times 10^{-16}$
$\pi_{SH} - \pi_{MNL}$	0.38	0.65

centile score estimates, respectively. For this dataset, the discrete choice models outperform the hot spot model with considerable statistical significance. Between the discrete choice models, there is not a significant difference in mean percentile score.

The improvement attained using discrete choice models on this dataset is consistent with results seen using other spatial attack datasets, including the dataset of suicide bombing incidents discussed in Brown et al. (2004).

5 Conclusion

This paper presented a newly developed algorithm for the objective application of multinomial discrete choice models to spatial point patterns. The algorithm was exercised using a data set of urban residential breaking and entering crimes. The evaluation results demonstrate that models produced by the algorithm can predictively outperform traditional models not based on discrete choice theory, and are often more interpretable by the analyst. Some important characteristics of the algorithm include:

- the use of correlation clustering in feature selection to minimize multicollinearity;
- the use of a separability metric in feature selection to maximize predictive performance;
- the use of cross validation to estimate performance and select appropriate model complexity;

The algorithm can be integrated into a spatial crime analysis system such as ReCAP (Brown 1998) to aid planners without requiring experience in discrete choice modeling.

References

Ben-Akiva M, Lerman SR (1985) Discrete choice analysis. The MIT Press, Cambridge

Brown DE (1998) The regional crime analysis program (ReCAP): a framework for mining data to catch criminals. In: Proceedings of the 1998 IEEE international conference on systems, man, and cybernetics, pp 2848–2853

Brown DE, Dalton J, Hoyle H (2004) Spatial forecast methods for terrorist events in urban environments. In: Proceedings of the second NSF/NIJ symposium on intelligence and security informatics. Lecture Notes in Computer Science, Tuscon, Arizona. Springer, Heidelberg, pp 426–435

Fotheringham AS (1986) Modelling hierarchical destination choice. Environ Plan A 18:401–418

Fotheringham AS, O'Kelly ME (1989) Spatial interaction models: formulations and applications. Kluwer, Dordrecht

Fotheringham AS, Brunsdon C, Charlton M (2000) Quantitative geography. Sage Publications Ltd, London

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning; data mining inference and prediction. Springer, Heidelberg

Johnson NLm, Kotz S (1970) Continuous Univariate Distributions—1. Wiley, New York

Lia H, Brown DE (2003) Criminal incident prediction using a a point-pattern-based density model. Int J Forecasting 19:603–622

McFadden D (1974) Conditional logit analysis of qualitative choice behavior, Chapter 4. Academic, New York, pp 105–142

Team RDC (2004) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna