VU Research Portal

The Importance of Importance Sampling

Curtis-Ham, Sophie; Bernasco, Wim; Medvedev, Oleg N.; Polaschek, Devon L.L.

Link to publication in VU Research Portal

ORIGINAL PAPER

The Importance of Importance Sampling: Exploring Methods of Sampling from Alternatives in Discrete Choice Models of Crime Location Choice

Sophie Curtis-Ham[1] · Wim Bernasco[2,3] · Oleg N. Medvedev[1] · Devon L. L. Polaschek[1]

Abstract

Objectives The burgeoning field of individual level crime location choice research has required increasingly large datasets to model complex relationships between the attributes of potential crime locations and offenders' choices. This study tests methods of sampling aiming to overcome computational challenges involved in the use of such large datasets.

Methods Using police data on 38,120 residential and non-residential burglary, commercial and personal robbery and extra-familial sex offense locations and the offenders' pre-offense activity locations (e.g., home, family members' homes and prior crime locations), and in the context of the conditional logit formulation of the discrete spatial choice model, we tested a novel method for importance sampling of alternatives. The method over-samples potential crime locations near to offenders' activity locations that are more likely to be chosen for crime. We compared variants of this method with simple random sampling.

Results Importance sampling produced results more consistent with those produced without sampling compared with simple random sampling, and provided considerable computational savings. There were strong relationships between the locations of offenders' prior criminal and non-criminal activities and their crime locations.

Conclusions Importance sampling from alternatives is a relatively simple and effective method that enables future studies to use larger datasets (e.g., with more variables, wider study areas, or more granular spatial or spatio-temporal units) to yield greater insights into crime location choice. By examining non-residential burglary and sexual offenses, in New Zealand, the substantive results represent a novel contribution to the growing literature on offenders' spatial decision making.

Keywords Crime location choice · Discrete choice modelling · Police data · Routine activity nodes · Sampling from alternatives

✉ Sophie Curtis-Ham
  SC398@students.waikato.ac.nz

Extended author information available on the last page of the article

Introduction

Understanding crime location choice at an individual level is a growing research undertaking with potential to inform criminal investigation and prevention activities beyond improvement in the explanation and prediction of where and when crime concentrates (Bernasco 2017; Ruiter 2017). There is an opportunity to deepen this understanding by increasing its geographic granularity. Whereas most prior studies of crime location choice focused on neighborhoods, recent research has demonstrated the relevance of smaller and more ecologically valid spatial units of analysis, such as streets segments (Frith et al. 2017), census blocks (Bernasco et al. 2013), and even individual properties (Vandeviver et al. 2015). Such smaller units can better account for heterogeneity in relevant variables that would be diluted within larger spatial units, so that the effects of these variables would become undetectable. In addition, because criminal opportunities are subject to cyclical temporal variations over days, weeks, months and seasons (Bernasco et al. 2017; van Sleeuwen et al. 2018), we need not only ask questions about optimal *places* for crime but also about optimal *times and places*, in combination, for crime. Further, there is a need to extend the scope or range of variables included in models of crime location choice. In particular, the measurement of individual activity spaces has been limited to residential addresses and prior offending locations only. Increasing the number of variables in crime location choice models can mean more offenses are needed in the dataset to enable sufficient statistical power.

However, because crime location choice studies combine data at the level of individual crimes with data aggregated to the spatial (or spatio-temporal) unit of analysis (e.g., neighborhoods, street segments or census blocks in four six-hour time blocks), the models that they utilize require estimation along three rather than two dimensions. Their estimation involves optimization along the dimensions of crimes×locations×variables, instead of crimes×variables or locations×variables. For example, a study with 1,000 crimes and 1,000 potential alternative locations for each crime, yields a dataset of 1 million rows. A study with more offenses, smaller units of analysis or a wider study area could involve upwards of 10,000 choices and 10,000 alternatives, yielding an unwieldy dataset of 100 million rows. If a temporal dimension were added, even as few as 1,000 spatial alternatives and 10 time periods would produce 10,000 space–time alternatives. These extensions require computer storage and processing capabilities that exceed the limits of available computing equipment outside of specialist computing labs (Vandeviver et al. 2015). Therefore, for model estimation to remain tractable outside of high-performance computer labs, there is a need to explore ways of reducing the computational burden. In fact, with the quick proliferation of model extensions and the advent of big data, high-performance computing environments may only offer a very short-term solution.

The issue of computational burden due to many choice alternatives has been addressed in other fields and on other types of decisions. Based on the results of McFadden (1977), researchers have estimated choice models by sampling from the decision makers' available choice sets when studying decisions such as residential choice: where to find/buy a house (Duncombe et al. 2001) and transportation route choice: how to travel from A to B (Frejinger et al. 2009). Within the discrete spatial choice modelling (DSCM) paradigm for studying crime location choice (see Ruiter 2017, for a review), only four studies (detailed below) have 'sampled from alternatives' instead of using the full set of locations that could have been chosen by any given offender. Specifically, these studies have used simple random sampling (McFadden 1977). However, research in other location choice domains suggests

that simple random sampling may not necessarily be an optimal strategy if a priori knowledge on choice probabilities is available from previous research. For example, importance sampling strategies, that over-sample from the a priori most likely alternatives to be chosen (McFadden 1977), can lead to estimates that are closer to those produced by including all alternatives (Lemp and Kockelman 2012; Hassan et al. 2019). We therefore examined the effects of different methods of sampling from alternatives on the results of discrete spatial choice using a real-world crime dataset. To our knowledge this is the first study to directly compare different sampling methods in the crime location choice context. It contributes to a burgeoning literature on crime location choice (Ruiter 2017) and provides initial indications of the effects of different sampling strategies which could help guide future studies in this paradigm. Although not the primary purpose of this paper, we also provide new insights into crime location choice from a country that has not yet featured in the DSCM literature.

We begin with a review of DSCM crime location choice studies with a focus on sample sizes and methods. We then discuss the literature on sampling from alternatives, which informed the selection of sampling strategies to compare in the present study. The experimental design and data used in this study are then described, along with the discrete choice model method used. Our results present the effects of sampling strategies on model coefficients and measures of fit, and we conclude by discussing their implications for sampling in future crime location choice studies.

Studies of crime location choice

A growing body of criminological research has sought to model offenders' decisions[1] about where to commit crime using a discrete choice approach. Discrete choice models (McFadden 1984) are common in other domains where decision makers are choosing from a range of alternatives—consumer products (Nevo 2001), transport modes (Nguyen et al. 2017), travel destinations (Huybers 2005)—and where researchers are interested in the attributes of the alternatives, and of the choosers, that influence the outcome of the decision. Studying crime location choice using discrete choice methods enables researchers to isolate attributes of offenders and potential crime locations that are associated with a location being chosen for crime (Townsley 2016; Bernasco 2017; Ruiter 2017). Attributes of locations (such as the number of potential crime targets present), of offenders (such as their age or level of criminal expertise) or of location-offender combinations (such as the location's distance from the offender's home), are input as predictor variables and the outcome variable is categorical: which of the possible locations was chosen for crime commission (e.g., Bernasco and Nieuwbeerta 2005; Long et al. 2018; Frith 2019)? To date, over thirty studies have applied discrete spatial choice modelling (DSCM) to crime data (see Ruiter, 2017 and subsequent studies e.g., Bernasco 2019; Hanayama et al. 2018; Long et al. 2018; Song et al. 2019). Their results have provided significant insights with implications for theory and practice in terms of crime investigation, prediction, and prevention (Bernasco 2019; Curtis-Ham et al. 2020) but there is potential to expand the research agenda further

---

[1] We use the terms 'decision' and 'choice' to refer to location choice as revealed in behaviour. The choice may not feel like a choice to the offender and may not even take place consciously. It can also reflect a decision to visit a place for a non-criminal purpose, whereupon a crime opportunity is identified and acted on (Ruiter 2017).

(Ruiter 2017; Curtis-Ham et al. 2020). Finding ways to overcome barriers to this expansion is therefore important.

In DSCM studies of crime location choice, the number of alternative locations that could be chosen for crime can be very large, as can the number of crime choices. Further, estimating discrete choice models requires attributes of all alternatives to be linked to the attributes of every choice. The datasets of all alternatives x all choices used (without sampling) so far have ranged from 5,502 rows (262 street segments×21 drug deals; Bernasco and Jacques 2015) to 93,919,959 rows (138,321 houses×679 burglaries; Vandeviver and Bernasco 2020). The more typical range is between approximately 500,000 to 5 million rows, involving 1000 to 5000 alternatives and 500 to 5000 offenses (Townsley et al. 2015; e.g., Bernasco et al. 2015; Frith et al. 2017). When these datasets outstrip the capacity of the typical research computer,[2] two options to reduce the dataset exist: reduce the number of choices, or the number of alternatives for each choice. However, the former may not be desirable. As with any form of regression, a smaller sample (here, of choices) means less statistical power, reducing the ability to detect associations between the attributes of alternatives and choices. Further, the more attributes being examined, the larger the sample (the more choices) needed to ensure sufficient power. We therefore focus on the latter option: sampling the alternatives.

Computational limitations have prompted the use of sampling from alternatives in four crime location choice DSCM studies to date. For example, Vandeviver et al's. (2015) dataset included 503,589 alternative houses that could be chosen by each of 650 residential burglars. Including all alternatives for each offender would yield 327,332,850 rows, beyond the processing capability of even the specialist computer lab used by the researchers. They therefore randomly sampled one of every 8 alternative addresses for each offender, yielding a manageable 40,916,200 rows (given the lab's computing capacity). Likewise, Bernasco et al. (2013) randomly sampled 5,999 from the 24,593 alternative census blocks that could have been chosen.[3] In addition, the authors randomly sampled 6,000 street robberies from 12,938 cases,[4] resulting in 36 million rows, processed on a consumer level computer with 12 GB of RAM. Promisingly, the estimates and standard errors were very close to those produced by models using the full choice set for a smaller sample of 2,000 offenders. In a further study with the same dataset Bernasco et al. (2017) randomly sampled 7,999 and 11,999 of the census blocks for models estimated using only the offenses occurring on a given day of the week and 2-h period of the day, respectively. Simple random sampling from alternatives was also used by Bernasco (2010) to reduce a potential dataset of almost 45 m rows to 2.8 m (sampling 1,499 of 23,984 alternative postcodes for 1871 burglaries).

To increase the robustness of the estimates by reducing the influence of random error introduced by sampling, an additional bootstrapping process was used by Vandeviver et al. (2015) and Bernasco et al. (2013). In these studies, model outputs were combined across 20 and 25 sampling iterations respectively. Although bootstrapping multiple samples iteratively can solve an issue of insufficient RAM to hold the full dataset, dividing a single long

---

[2] A dataset might be too large to hold in RAM, or it might take days, weeks or even months to run the model, depending on the speed of the processor.

[3] Note that the chosen alternative is always included in the sample; the random sample is taken from the remaining alternatives (McFadden 1977; Ben-Akiva and Lerman 1985).

[4] The fact that both sample sizes (of alternatives and of robberies) totalled 6,000 is coincidental. There is no reason why this should be the case.

processing time into many shorter processing times may not solve a problem of insufficient CPU power (processing speed).

However, simple random sampling may not adequately capture the alternatives considered by individual decision makers that contain the most information about variables relevant to their choice. In crime location choice it is very unlikely that offenders consider every possible alternative (at house, street or neighborhood level) equally carefully when deciding where to commit crime (Ruiter 2017). They are most likely to consider places of which they have existing knowledge (Brantingham and Brantingham 1991; Ruiter 2017; Menting 2018). To the analyst, locations known to offenders through their routine activities thus contain the most 'signal' about the variables influencing offenders' crime location choices, such as how well they know the location and its crime potential (Curtis-Ham et al. 2020). But the larger the number of alternatives, and the smaller the proportion of alternatives sampled, the less likely a random sample is to include these more informative alternatives. Thus, for example, the consistency between estimates from simple random sampling of 24% of alternatives and from the full choice set achieved by Bernasco et al. (2013) might not generalize to smaller samples. We next consider other means of sampling from alternatives that could potentially provide more robust estimates if applied to conditional logit models of crime location choice through prioritizing the sampling of the most informative alternatives. We focus on conditional logit because it is by far the most used model in in DSCM studies to date, and also because the development of sampling from alternatives in other discrete choice models is not yet completely developed (Guevara and Ben-Akiva 2013a, b; Guevara et al. 2016).

Alternative methods of sampling from alternatives

An alternative to simple random sampling for DSCMs is importance sampling, where alternatives that are more likely—based on a priori beliefs—to be chosen are preferentially sampled (McFadden 1977; Ben-Akiva and Lerman 1985). Importance samples are typically stratified: alternatives most likely to be chosen are sampled at a higher rate, followed by alternatives with lower (a priori) choice probabilities, for a number of strata defined by the researchers (Li et al. 2005). Methods of importance sampling range in complexity. Most simply, one could have a single stratum sampled randomly at a higher rate than the remainder. For example, Bhat et al. (1998) defined a single 'most feasible choice set' stratum as potential travel destinations that were within the maximum distance travelled in any of the trips in the dataset (see similarly, Shiftan 1998). More complex methods have defined multiple strata using a combination of their spatial location (often with reference to journey start points) and other theoretically relevant attributes. For example, in a study of residential location choice, zones located within a central city area (more desirable) and in the same income bracket as decision makers' current home zones were preferentially sampled (Bowman and Ben-Akiva 2001; see similarly Jonnalagadda et al. 2001). Several studies have used Moran's I to identify, statistically, strata made up of zones that are both spatially proximal and similar on a relevant variable such as the number of employees, when modelling work trip destinations (Li et al. 2005; Park et al. 2013; Kim and Lee 2017).

More sophisticated methods to establish the prior probability of each alternative being chosen, to inform its sampling probability, have been proposed. These include: using fuzzy logic to identify the routes (in a route choice scenario) most likely to be considered by individuals (Hassan et al. 2019); and using the choice probabilities output by an initial random sample (Lemp and Kockelman 2012). Moreover, when compared with simple

random sampling, such importance sampling methods lead to more robust results, producing smaller standard errors and higher predictive accuracy (Lemp and Kockelman 2012; Hassan et al. 2019).

Present study

We therefore propose a method for importance sampling in the crime location choice context and compare variants of this method with simple random sampling from alternatives. We also explore the impact of sample size, given previous demonstrations of the effects of sample size on model performance with both random (Nerella and Bhat 2004; von Haefen and Domanski 2013) and importance sampling (Park et al. 2013; Hassan et al. 2019). We employ a simple method for determining importance sampling strata akin to the use of distance from origin point in studies of trip destination choice (Bhat et al. 1998; Shiftan 1998). However, in crime location choice, the focus is increasingly on the relationship between multiple routine activity locations—the various locations frequented in daily life such as home, work, school, shops and family and friends' homes—and crime locations, rather than the origin and destination point of a specific journey (Ruiter 2017; Bernasco 2019; Menting et al. 2020). We also have theoretical and empirical grounds for believing that alternatives close to these routine activity 'nodes' are more likely to be chosen for crime commission than other alternatives (Brantingham and Brantingham 1991; Ruiter 2017; Bernasco 2019; Menting et al. 2020). The sampling method thus prioritizes the inclusion of alternatives closer to *any* of the routine activity nodes in the dataset as more likely to be in a given individual's choice set. To determine which sampling procedures best approximate the true estimates, we also run models using the full set of alternatives as a baseline. We separately study 5 different crime types, to account for the potential for different spatial relationships to exist for different crime types (Curtis-Ham et al. 2020) and to enable assessment of whether the results from different sampling methods generalize across crime types.

Method

Offense and offender data

The data used in this study included solved residential and non-residential burglary, commercial and personal robbery and extra-familial sex offenses occurring between 2009 and 2018 from a national dataset obtained from the New Zealand Police. For each of the offenders recorded as having committed these offenses, the location of their most recent offense was the location choice of interest. The dataset also included the locations of a range of pre-offense activity nodes. These included: past and present homes of the offender and their family members, school and other educational institutions attended, workplace, prior offenses they had committed or experienced as victims or witnesses, non-crime incidents in which they were involved, and places they were arrested, stopped or otherwise noted for intelligence purposes. Curtis-Ham et al. (2021) describe the dataset in detail; it

includes approximately 4.5 million activity locations for 60,607 offenders. In this study we used random samples of 50% of the offenders within each offense type.[5]

Because these activity node locations are only recorded where needed for operational purposes during policing activities, they are not a complete or systematic set of pre-offense activity locations for each offender. However, they constitute a wider array of activity nodes than used in previous crime location studies based on administrative data (e.g., Lammers et al. 2015; Menting et al. 2016; van Sleeuwen et al. 2018). Further, Curtis-Ham et al. (2021) analyze the extent—number and geographic range-of offenders' pre-offense activity nodes in this dataset, concluding that the data hold potential for use in crime location choice research. Indeed, the results of the present study confirm that the activity nodes included in the dataset provide considerable 'signal' in explaining offenders' crime locations.

Unit of analysis

In this study we used the NZ Census Statistical Area 2 (SA2) as the set of locations from which an offense location could be chosen. SA2s roughly equate to neighborhoods and typically contain 2000–4000 residents in metropolitan areas (1000–5000 in rural areas). There were 2153 SA2s, with land areas of $0.063km^2$ to $12,042.36km^2$, reflecting the relative population density in urban and rural areas (median $1.962km^2$).[6] SA2s are comparable to the units used in other crime location choice studies (e.g., Clare et al. 2009; Townsley et al. 2015).

Outcome variable

The outcome variable was the choice of SA2: in which of the 2153 SA2 areas of New Zealand did the offender commit the index offense? Whereas in the *theoretical model* we assume that all SA2s appear in the offender's choice set, in the *estimation* of the parameters of this model for each offense only a subset of the SA2s is used.

Predictor variables

As our focus was on testing different sampling methods, rather than testing detailed explanatory factors, we constructed a simple model using six predictors reflecting the proximity of offenders' routine activity nodes to each SA2 in their choice set, and an additional seventh predictor reflecting the level of crime opportunity in each SA2. Previous crime location choice studies have demonstrated that the odds of a neighborhood being chosen for crime are greatest when there is an activity node in the same neighborhood and lower for neighborhoods that are further from any of the offender's activity nodes (Bernasco 2019;

---

[5] This study forms part of a wider programme of research for which the data were divided into 'training' and 'test' samples (50% each). The training data were used for all analyses where models were trained (such as the present study). The test data were reserved for later studies testing model accuracy when applied to new data.

[6] The 2018 SA2 shapefile and metadata were downloaded from https://datafinder.stats.govt.nz/layer/92212-statistical-area-2-2018-generalised/. We excluded 83 SA2s which cover large bodies of water along coastlines and over lakes.

Menting et al. 2020). We therefore included dichotomous variables reflecting the presence or absence of offenders' activity nodes in increasing distance bands in relation to each SA2 in their choice set. The six distance bands were: within the same SA2, or within 0-200 m, 200-500 m, 500 m-1 km, 1-2 km or 2-5 km outside of the SA2 boundary. Following previous crime location choice research (e.g., Menting et al. 2020), the activity node variables reflected whether the *nearest* activity node to the SA2 fell into a given distance band. For example, if the closest node was in the same neighborhood, we coded 'Same SA2' as true and all other node distance variables false; if the closest node was 4 km outside the SA2 boundary, we coded 'Node within 2-5 km' as true and all other node distance variables false.

Since crime location choice is not only a product of the locations of which offenders are aware from their routine activities but of the opportunities available at those locations (Brantingham and Brantingham 1991; Menting 2018), we also included a measure of opportunity relevant to each crime type. The opportunity measures were sourced from Statistics NZ Census and Business Demography data (http://nzdotstat.stats.govt.nz/) and are comparable to opportunity measures used in other crime location choice studies (Townsley et al. 2015; Lammers 2018; Long et al. 2018; e.g., Frith 2019). For residential burglary, opportunity was the number of households in the SA2.[7] For non-residential burglary it was the number of business units in any industry.[8] For commercial robbery, it was the number of business units for industries of the types targeted in commercial robbery.[9] For personal robbery and extra-familial sex offenses it was the number of commercial or public business units, as an indicator of ambient population and thus the number of potential crime targets.[10]

Sampling strategies

We compared nine strategies for sampling from alternatives to the results from the full set of alternatives. The sampling strategies are described in Table 1 and form three groups, within which we varied the sample size. The first group involved 'distance importance sampling' (DIS) strategies where we included all SA2s within 5 km of the offender's activity nodes and added more strata at increasing distances from which alternatives were randomly sampled with decreasing probability. Similar distance-based importance strata were used in previous non-crime location choices studies (Ben-Akiva and Bowman 1998; Shiftan 1998; Li et al. 2005). In the absence of evidence as to how many strata to include over what distance in studying crime location choice, we included three additional distance strata

---

[7] There were large changes in residential population in many SA2s over the data period due to the Christchurch earthquakes and housing developments in response to increasing urban populations. Census 2013 data was used for offenses occurring between 2009 and 2015, and census 2018 data were used for offenses occurring between 2016 and 2018.

[8] Business demography statistics remained consistent over the data period so 2018 was used for simplicity.

[9] Industry categories included: G Retail Trade, H Accommodation and Food Services, K Financial and Insurance Services, L Rental, Hiring and Real Estate Services. M Professional, Scientific and Technical Services, K Financial and Insurance Services, L Rental, Hiring and Real Estate Services, M Professional, Scientific and Technical Services, N Administrative and Support Services, R Arts and Recreation Services, S Other Services. See Curtis-Ham et al. (2021) for details of how 'commercial' robberies were identified.

[10] All industries as for commercial robbery plus: I Transport, Postal and Warehousing, J Information Media and Telecommunications, O Public Administration and Safety, P Education and Training, Q Health Care and Social Assistance.

Table 1 Strategies used for sampling from alternatives

Label	Sample includes chosen SA2 plus:	Sample size equivalence[a]
DIS1	Stratum: 1: SA2s that contained or had activity nodes within 5 km of the SA2 boundary	*
	Remainder stratum: 10 randomly sampled from the remaining SA2s outside of other strata	
DIS2	As above plus:	**
	Stratum 2: 20 sampled from remaining SA2s between 5 and 10 km of any activity node.[b]	
DIS3	As above plus:	
	Stratum 3: 15 sampled from remaining SA2s between 10 and 50 km of any activity node.[b]	
DIS4	As above plus:	***
	Stratum 4: 10 sampled from remaining SA2s between 50 and 100 km of any activity node.[b]	
SIS1	Stratum 1 plus 30 sampled from remaining SA2s	**
SIS2	Stratum 1 plus 55 sampled from remaining SA2s	***
SIS3	Stratum 1 plus 100 sampled from remaining SA2s	****
SRS1	Random sample of 137 to 261 (depending on crime type)	*
SRS2	Random sample of 227 to 345 (depending on crime type)	****

[a]Asterisks indicate strategies with comparable sample sizes

[b]Identified as SA2s whose centroids were within the specified distance range from the centroid of any SA2 containing an activity node. If there were fewer than the specified number of SA2s available to sample, 100% were sampled

beyond the initial 0-5 km (strategy DIS1): 5-10 km (DIS2), 10-50 km (DIS3), 50-100 km (DIS4). These enabled us to investigate the incremental benefit (if any) of importance sampling SA2 alternatives at a series of increasing distances.

The second group, 'simple importance sampling' (SIS), included all SA2s within 5 km of the offender's activity nodes and increased the number of additional SA2s sampled from the remainder. The simple importance strategies were included to test whether any advantage of the distance importance sampling was attributable to the inclusion of SA2s in the distance strata or merely to the inclusion of additional SA2s regardless of their location.

The third group included two simple random sampling (SRS) strategies based on the smallest and largest sample sizes achieved with the previous strategies (to enable comparison of like for like in terms of sample size). For the first SRS strategy (SRS1) we randomly sampled the median number of SA2s included in the choice sets when using the first distance importance sampling strategy (DIS1). This strategy resulted in the *smallest* choice sets of all the strategies (see Table 2 below). For the second SRS strategy (SRS2) we randomly sampled the median number of SA2s included in the choice sets when using the sampling strategy that resulted in the *largest* choice sets (SIS3: all SA2s within 5 km of activity nodes plus 100 from the remainder, see Table 2).

The number of additional SA2s to sample were determined with reference to previous research suggesting that robust estimates could be achieved by randomly sampling 12.5% of the full set of alternatives (Nerella and Bhat 2004) and studies using stratified importance samples as small as 1–7% (Bowman and Ben-Akiva 2001; Jonnalagadda et al. 2001;

Table 2 Choice set size, total dataset size and run time per sampling strategy and offense

Offense	Sampling strategy & size equivalence (*)[a]	Total N in dataset	% of All alts N[b]	n alternatives per choice set			Run time (s)	
				Min	Median	Max	Mean	SD
Res. Burg. (n = 17,054)	DIS1	*	4,213,998	11.5%	11	210	1252	42.4	1.3
	DIS2	**	4,501,322	12.3%	11	229	1272	45.8	0.6
	DIS3		4,754,811	12.9%	11	244	1287	48.9	1.1
	DIS4	***	4,925,329	13.4%	20	254	1297	50.7	1.1
	SIS1	**	4,555,078	12.4%	31	230	1272	46.2	0.9
	SIS2	***	4,981,428	13.6%	56	255	1297	51.1	0.9
	SIS3	****	5,748,858	15.7%	101	300	1342	58.8	0.9
	SRS1	*	3,598,394	9.8%	210	210	210	35.0	0.8
	SRS2	****	5,133,254	14.0%	300	300	300	50.5	0.8
	All alts		36,717,262	100.0%	2153	2153	2153	649.1	386.7
Non. Res. Burg. (n = 10,353)	DIS1	*	2,417,735	10.8%	11	190	1114	25.3	1.0
	DIS2	**	2,587,458	11.6%	11	209	1134	26.1	0.8
	DIS3		2,740,966	12.3%	12	224	1149	28.2	0.8
	DIS4	***	2,844,471	12.8%	20	234	1159	29.4	0.8
	SIS1	**	2,624,795	11.8%	31	210	1134	27.0	0.6
	SIS2	***	2,883,620	12.9%	56	235	1159	29.8	0.9
	SIS3	****	3,349,505	15.0%	101	280	1204	34.1	0.7
	SRS1	*	1,977,423	8.9%	190	190	190	18.9	0.5
	SRS2	****	2,909,193	13.1%	280	280	280	28.5	0.9
	All alts		22,290,009	100.0%	2153	2153	2153	244.8	5.9

Table 2 (continued)

Com. Rob.(n=1,977)								
DIS1	*	563,161	13.2%	13	261	1028	5.7	0.6
DIS2	**	598,905	14.1%	13	281	1048	5.8	0.2
DIS3		628,470	14.8%	28	296	1063	6.3	0.4
DIS4	***	648,240	15.2%	38	306	1073	6.5	0.3
SIS1	**	602,701	14.2%	33	281	1048	5.9	0.2
SIS2	***	652,126	15.3%	58	306	1073	6.7	0.6
SIS3	****	741,091	17.4%	103	351	1118	7.4	0.3
SRS1	*	517,974	12.2%	261	261	261	4.8	0.2
SRS2	*****	695,904	16.3%	351	351	351	6.8	0.3
All alts		4,256,481	100.0%	2153	2153	2153	40.5	1.9
Pers. Rob.(n=4,315)								
DIS1	*	1,243,204	13.4%	13	255	1067	13.0	0.6
DIS2	**	1,321,381	14.2%	13	275	1087	13.6	0.6
DIS3		1,385,822	14.9%	13	290	1102	14.4	0.6
DIS4	***	1,428,972	15.4%	23	300	1112	15.9	2.6
SIS1	**	1,329,504	14.3%	33	275	1087	13.6	0.8
SIS2	***	1,437,379	15.5%	58	300	1112	14.5	0.5
SIS3	*****	1,631,554	17.6%	103	345	1157	16.6	0.8
SRS1	*	1,104,640	11.9%	255	255	255	10.8	0.3
SRS2	*****	1,492,990	16.1%	345	345	345	15.0	0.4
All alts		9,290,195	100.0%	2153	2153	2153	91.4	1.3

Table 2 (continued)

Sex offense (n = 4,421)								
DIS1	*	784,537	8.2%	11	137	1113	8.8	2.3
DIS2	**	854,011	9.0%	11	156	1133	8.5	1.1
DIS3		919,350	9.7%	14	171	1148	8.5	0.4
DIS4	***	963,554	10.1%	24	181	1158	9.3	0.6
SIS1	**	872,957	9.2%	31	157	1133	8.0	0.1
SIS2	***	983,482	10.3%	56	182	1158	9.1	0.2
SIS3	****	1,182,427	12.4%	101	227	1203	12.2	0.5
SRS1	*	610,098	6.4%	137	137	137	5.9	0.3
SRS2	****	1,007,988	10.6%	227	227	227	10.2	0.3
All alt		9,518,413	100.0%	2153	2153	2153	92.2	1.7

[a] Asterisks indicate strategies with comparable sample sizes

[b] Total N of the sampled dataset as a percentage of the Total N of the unsampled dataset using all alternatives

Li et al. 2005; Kim and Lee 2017). While the size of the choice set varies across individuals in the present study, the sampling strategies aimed to generate choice sets averaging between 5 and 15% of the 2153 SA2s (about 100–320 SA2s).

Model specification and estimation

The *conditional logit model* (McFadden 1974) is a statistical model for the probability that a decision maker *n*, who must choose from a set of alternatives *C*, chooses alternative *i*, and can be expressed as:

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_{j \int C} e^{\beta' x_{ni}}} \tag{1}$$

where x_{ni} is a list of attributes that vary across alternatives and may also vary across decision makers, and ß is a vector of the parameters that represent the effects of these attributes on the outcome of the decision. The ß parameters can be estimated by maximum likelihood estimation, based on the actual choices observed. From the size, direction and statistical significance of the estimated β parameters, conclusions can be drawn about the relevant criteria that decision makers use. Typically, exponentiated ß estimates (e^{β}) are reported. They are called odds ratios and represent the effect of a one-unit increase in the x_{ni} variables on the odds of an alternative to be chosen.

The log-likelihood of the conditional logit model is:

$$l = \sum_{n=1}^{N} \sum_{i \int C, i \neq j} \left(y_{ni} \ln \left(P_{ni} \right) \right) = \sum_{n=1}^{N} \sum_{i \int C, i \neq j} \left(y_{ni} \ln \left(\frac{e^{\beta' x_{ni}}}{\sum_{j \int C} e^{\beta' x_{nj}}} \right) \right) \tag{2}$$

where y_{in} is the observed choice, such that $y_{ni} = 1$ if decision maker *n* chooses alternative *i* and $y_{ni} = 0$ if another alternative is chosen.

Sampling from alternatives is an estimation technique in which we use a subset *D* of the full choice set *C* to estimate the ß parameters. McFadden (1977) proved that under the *positive conditioning property* (whereby each alternative in the choice set *C* has a positive probability of being included in the estimation set *D*), unbiased parameter estimates are consistently estimated by maximizing a modified likelihood function with an added correction term $-ln(\pi(D|i))$ in the utility function:

$$\ell = \sum_{n=1}^{N} \sum_{i \int C, i \neq j} \left(y_{ni} \ln \left(\frac{e^{\beta' x_{ni} - \ln(\pi(D|i))}}{\sum_{j \int C} e^{\beta' x_{nj} - \ln(\pi(D|j))}} \right) \right) \tag{3}$$

where $\pi(D|i)$ is the probability of alternative *i* to be included in the estimation sample *D*. This modified estimation procedure is quite general, as it applies to *any* sample that conforms to the positive conditional property. All *importance sampling* strategies that we use in the present paper conform to the positive conditioning property, and were therefore estimated with the additional offset term. In all cases, the probability of the alternative being sampled depended on the number of alternatives remaining to be sampled after including those with activity nodes within 5 km.

In the case of a *uniform conditional probability*, as in the case of *simple random sampling*, each alternative from the full choice set *C* has the same positive probability of being included in D. The $\pi(D|i)$ thus cancels out in Eq. (3), and the model parameters can be estimated by the regular log-likelihood Eq. (2).
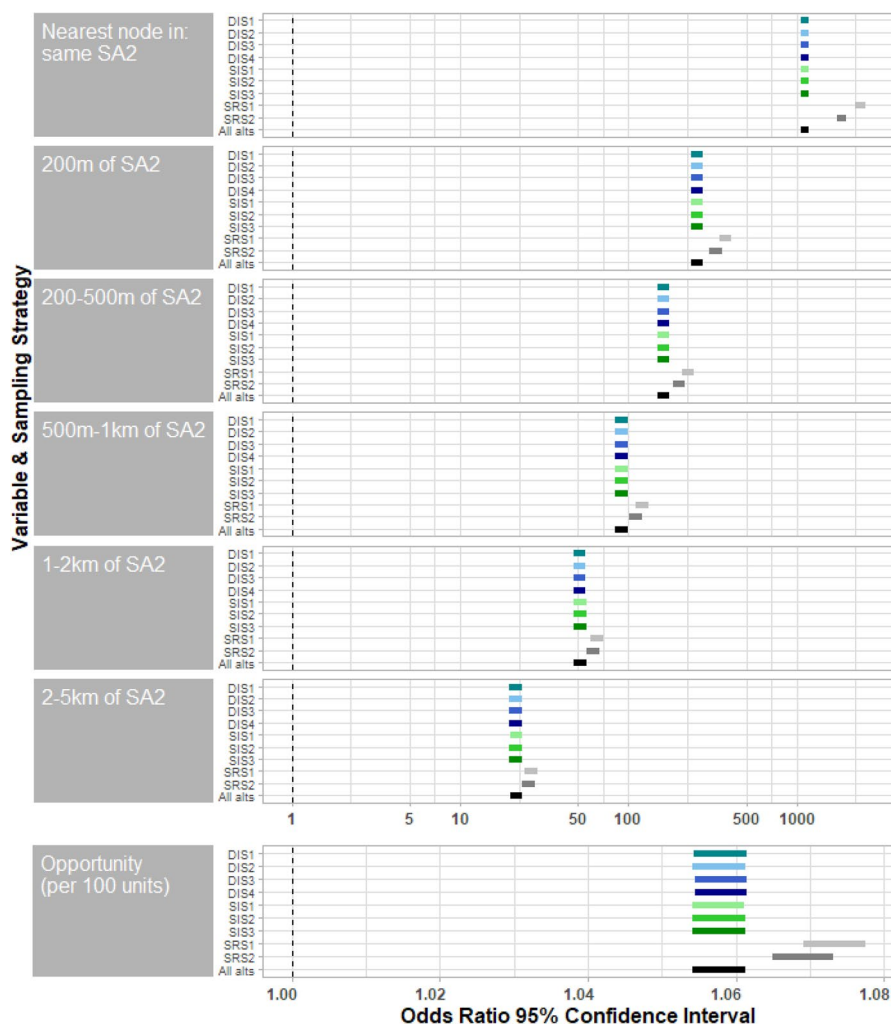
Fig. 1 Parameter estimates for residential burglary location choice by sampling strategy

All models were run on an HP Elitebook with 16gb of RAM and an Intel i7-7600U 2.8 GHz CPU using the clogit function from the survival package (Therneau 2020) in R (R Core Team 2013).

Comparison metrics

We assessed the quality of the sampling strategies with a range of criteria. First, we consider the size of the dataset produced by each strategy, the aim being to minimize the number of rows and thus the computational burden while producing robust results. We also compare the processing times for the models based on each strategy as an indicator of computational efficiency (over 10 iterations, using the microbenchmark package: Mersman 2019). The times are

of course specific to the computer used to run the analyses, but provide a general indication of the relative time saving of the sampling methods by comparison to using the full choice set. Second, we examine whether the strategies produce the same results as those from baseline model including the full choice set. Following previous studies comparing alternative-sampling methods (Park et al. 2013; Hassan et al. 2019), we consider the parameter estimates (Odds Ratios and their 95% confidence intervals) and model fit (McFadden's pseudo-R^2). In discussing the results of these measures, we also consider the complexity of the strategy, with simpler strategies that produce robust results being preferred. The R script used to sample the data, run models, and analyze the results is appended as Supplementary Material.

Results

We first present the results of the sampling strategies in terms of the overall dataset sizes and the distributions of the size of the choice set per offender, before comparing model estimates and fit.

Sample sizes and sampling probabilities

Table 2 shows for each offense and sampling strategy: the number of offenders (choices), the total number of rows in the dataset, the proportion of the 'all alternatives' dataset included in the sampled dataset, the minimum, median and maximum number of alternatives included in the choice set per offender, and the model processing times (average over 10 iterations). Within the importance sampling strategies, the size of choice sets varied widely, reflecting the variability in the number of activity nodes per offender in the dataset. The importance sampling strategies resulted in as few as 11 alternatives in the choice set (0.5% of the 2153 SA2s). But on average, offenders' choice sets contained at least 150–350 alternatives, representing roughly 7–16% of the 2153 SA2s, depending on crime type and sampling strategy. All of the sampling methods considerably reduced the total size of the datasets and therefore the computational burden, with the 'random minimum' strategy producing the smallest datasets. The proportions of SA2s sampled per *stratum* are provided in Table S1 in the Supplementary Materials. Run times reflected the size of the datasets, with sampling from alternatives models running in 5–18% of the time taken to run the all-alternatives models. Further, the computational savings were greatest for the datasets with more offenses (e.g., 5–9% for residential burglary), suggesting that the bigger the offense sample (number of choices) the greater the gains from sampling from alternatives.

Parameter estimates

Figure 1 displays the parameter estimates, by sampling strategy, for the residential burglary offenses (see Figs. 2–5 in Appendices 1 to 4 for other offenses). For the activity node variables, the odds ratios (ORs) represent the increase in probability of an SA2 being chosen for crime given the presence of an activity node in the given distance band. For example, offenders were over 1000 times more likely to commit a residential burglary in an SA2 in which they also had an activity node. These odds decreased over increasing distances to the nearest activity node but remained statistically significant; the presence of an activity node up to 5 km from an SA2 was associated with an over 20-fold increase in the likelihood that

Table 3 McFadden Pseudo R^2 values by sampling strategy and crime type

Sampling strategy & size[a]		Res. Burg	Non-res. Burg	Com. Rob	Pers. Rob	Sex Offense
DIS1	*	0.386	0.395	0.316	0.346	0.374
DIS2	**	0.386	0.395	0.316	0.346	0.374
DIS3		0.386	0.395	0.316	0.346	0.374
DIS4	***	0.386	0.395	0.316	0.346	0.374
SIS1	**	0.386	0.395	0.316	0.346	0.374
SIS2	***	0.386	0.395	0.316	0.346	0.374
SIS3	****	0.386	0.395	0.316	0.346	0.374
SRS1	*	0.404	0.421	0.323	0.358	0.418
SRS2	****	0.398	0.411	0.320	0.354	0.400
All alternatives		0.386	0.395	0.316	0.346	0.374

[a]Asterisks indicate strategies with comparable sample sizes

the SA2 would be chosen. The odds of an SA2 being chosen for a residential burglary also increased by 1.06 times for every 100 households in the SA2.

When comparing the sampling strategies with the 'all alternatives' model that includes all 2153 SA2s in each offender's choice set, all importance sampling strategies resulted in parameter estimates and standard errors that did not differ significantly from the full model. None of the incremental additions of the three distance strata (DIS2, DIS3, DIS4) beyond the initial 5 km (DIS1) provided additional benefits in terms of correspondence to the full model. Nor did increasing the sample size (SIS1, SIS2, SIS3) beyond the minimum achieved by including all SA2s with activity nodes within 5 km and 10 SA2s from the remainder (DIS1). However, the two simple random sampling strategies (SRS1 and SRS2) tended to produce coefficients that deviated widely from the full model coefficients (and larger standard errors). Deviation from the full model was larger for the variables reflecting the presence of activity nodes in closer proximity to the SA2, and for the smaller of the two random samples (SRS1). Conversely, the ORs for variables reflecting the presence of activity nodes farther from the SA2 were closer to those produced by the full model. The same broad pattern was found for the other offense types, though by comparison with burglary, for personal robbery and sex offenses the simple random sampling strategies (SRS1 and SRS2) produced ORs closer to those of the full model for variables reflecting activity nodes at longer distances from the SA2.[11] For commercial robbery, the confidence intervals produced by simple random sampling (SRS1 and SRS2) overlapped with those of the full model for all variables.

---

[11] We also compared bootstrapped versions of the single stratum importance sampling strategy (DIS1) and the smallest simple random sampling (SRS1), since the 'strategy to beat' to produce robust results with the smallest dataset was the single stratum importance sample, to which the smaller simple random sampling strategy was closest in sample size. The estimates and standard errors for 20 bootstrap iterations were combined using Rubin's rule (Rubin 1987) implemented in the Amelia package in R (King et al. 2000). The bootstrapped strategies produced the same pattern as the single iterations, as shown in Fig. 6 in Appendix 5. Of note, bootstrapping the simple random sampling did not produce estimates any closer to those from the full model.

Model fit

Table 3 presents the McFadden's pseudo r-squared values per sampling strategy and the full model, for each offense type. As with the parameter estimates, importance sampling strategies (DIS1-4 and SIS1-3) led to pseudo r-squared values that matched the values from the full model using all alternatives in each choice set. The simple random strategies led to slightly higher values than the full model, with the smaller random samples (SRS1) deviating the most. That smaller random samples led to higher R^2 values is consistent with previous studies (Park et al. 2013; Hassan et al. 2019).

Discussion

This study examined the effects of different methods of sampling from alternatives on the results of discrete spatial choice models of offenders' choice of crime locations, for burglary, robbery and extra-familial sex offenses. Our results suggest that overall, importance sampling that ensures the inclusion of choice alternatives near to offenders' activity nodes can lead to coefficients and model fit on par with the results of the full model using all alternatives in each choice set, while reducing the computational burden considerably. Simple random sampling, however, tends to risk overestimating both parameter estimates and model fit. Since all importance sampling strategies produced comparable results, considering both the size of the dataset and the complexity of the strategy, the single stratum based strategy (all SA2s with activity nodes within 5 km plus 10 SA2s from the remainder, DIS1) was the optimal strategy to produce robust results with the smallest dataset, fastest run time and simplest method. That preferentially sampling choice alternatives with higher choice probability outperforms simple random sampling is consistent with previous studies in other discrete spatial choice domains (Lemp and Kockelman 2012; Hassan et al. 2019).

In the only other study to compare sampling from alternatives to the full choice set in a crime location choice context, Bernasco et al. (2013) examined street robberies in the city of Chicago. They found that the coefficients and standard errors from a model using a simple random sample of 24% of 24,593 census block alternatives for 6000 robberies were very close to those from a model using the full choice set for 2000 robberies. There are several possible explanations for simple random sampling producing robust results in that context by comparison to the present study. First, and likely foremost, the proportion of alternatives sampled was considerably larger. Second, offenders may be familiar with a higher proportion of locations (alternatives) within a city than within an entire country; thus there would be a higher chance of sampling alternatives relevant to offenders' decisions. Third, by including only offenders with residential addresses in the city, who were probably familiar with more parts of the city than outsiders, their study likely includes more offenders with greater familiarity with more alternatives than the present study. In contrast, when using a national dataset, simple random sampling may not capture enough alternatives containing or near to activity nodes, to adequately capture the 'signal' from those alternatives that fall within offenders' awareness space (Brantingham and Brantingham 1991).

A further explanation relates to the size of the units of analysis. Bernasco et al. (2013) used small spatial units—census blocks with an average of 118 residents—by comparison to thousands in our neighborhood sized SA2s. It may be that simple random sampling performs poorly when sampling neighborhoods because features relevant to offenders' location choices (e.g., activity nodes, targets) concentrate in few neighbourhoods. Thus random

sampling may have a higher risk of excluding all of them. Conversely, these features may be present in more census blocks and thus the risk of excluding them all from the analysis by selecting areas randomly may be smaller.[12]

Overall, our results suggest that (a) simple random sampling does not necessarily lead to robust results and (b) ensuring that the sampling from alternatives strategy captures enough alternatives within individual offenders' awareness space is an important consideration when designing discrete crime location choice research. Particularly in countries or regions with high levels of inter-city mobility among offending populations, such as New Zealand (Curtis-Ham et al. 2021) and some European countries (Menting et al. 2020; Polišenská, 2008, as cited in Vandeviver et al. 2015; van Daele et al. 2012; van Daele and Vander Beken 2010) or parts of the USA (Bichler et al. 2012), studies of crime location choice may benefit from a wider focus. More widely focused studies employing neighborhood level units would also likely benefit from importance sampling.

The relative benefits of importance sampling also depended on crime type. Simple random sampling for commercial robbery produced results more consistent with the full model than for other offenses. A range of factors could explain this finding, likely in combination. First, commercial robbery has more specific targets so offenders may need to seek opportunities that are outside their activity space (at least as revealed by the present data). The prior choice probabilities for SA2s may thus be more evenly distributed within and outside the activity space limit (5 km) such that importance sampling and simple random sampling achieve more similar results. Second, robberies are more likely to involve co-offenders (Bright et al. 2020). With group offending, the awareness space of a single offender has less influence on crime location choice (Bernasco 2006; Lammers 2018), which would similarly lessen the difference between importance and simple random sampling. Third, commercial robbery offenders had more activity nodes on average than other offenders, meaning larger proportions of alternatives were sampled in both the importance and simple random sampling strategies. Lastly, commercial robbery had the smallest sample of offenders. The CIs are thus wider than for other offenses and wider CIs mean more potential for overlap between the random sample CIs and the full model CIs. If the relatively better performance of simple random sampling for this small sample of offenders were solely attributable to the CIs, it would be preferable to use importance sampling with future small offender samples, given it yielded results in line with the full model across the range of offender sample sizes covered by our different crime types.

The results for the different crime types suggest that importance sampling would be more important for crime types not included in this study that have more in common with burglary, personal robbery or sex offenses than commercial robbery. For example, property crimes where the targets are relatively ubiquitous, such as shoplifting, thefts of and from cars, and thefts from the person are more comparable to non-residential burglary or personal robbery and thus would likely benefit from importance sampling. However, predatory offenses targeting specific victim populations that require offenders to seek opportunities outside, or bearing less relation to their personal activity locations (e.g., sexual or other violent offenses targeting prostitutes in red light districts: Rossmo 2000) may be more akin to commercial robbery, with less need to over-sample alternatives in offenders' awareness space.

Our substantive findings as to the strong association between prior activity locations and crime location choice are also of significance. Even a simple model based on the presence (or not) of a range of activity nodes within a range of distances to a potential crime location, and the level of opportunity in that location, explained a substantial amount of

---

[12] We are grateful to an anonymous reviewer for contributing this explanation.

variance in crime location choice. These results are consistent with criminological theory (Brantingham and Brantingham 1991) and prior crime location choice studies that also used a range of activity nodes and found higher odds of crimes near offenders' activity nodes, declining with distance (Bernasco 2019; Menting et al. 2020). Our results are, however, novel in several respects. To our knowledge no prior crime location choice study has examined non-residential burglary or extra-familial sex offenses separately from other crimes. Our results confirm that activity space proximity is strongly related to crime location choice for these offenses.

Further, existing studies that included a comparably wide array of activity nodes have only measured their relationship to crime in general (Bernasco 2019; Menting et al. 2020), which may mask crime-type specific patterns. By disaggregating crime types, the present study revealed that while the overall trend of decreasing choice probability over increasing distance from activity nodes applies to each crime type studied, some notable variation exists. For example, commercial robbery displayed the smallest odds of crime location choice in close proximity to activity nodes. Commercial robbery tends to involve specific types of premises (e.g., convenience stores and petrol/gas stations) and offenders may need to search further afield to find targets that are not just available but suitable, considering for example their level of security, layout, and ease of escape (Taylor 2002; Altizio and York 2007). The ORs for sex offenses were closer to those of burglars than robbery offenders, with particularly high odds (~ 1000x) of crime in SA2s in which they had an activity node. These high odds may be partly explained by the inclusion of sex offenses that took place at the offender's home address. We note that sex offenders are a heterogenous group, with the present cohort including offenses against both adults and children, and known and stranger victims. These subgroups may have stronger or weaker associations between their home or other activity locations and crime locations, but victim information was not in the data to enable further disaggregation.

Some caveats apply to the present findings on the advantages of importance sampling from alternatives in the crime location choice context. The findings are based on a single study in one country, from one data source, requiring replication with datasets from other jurisdictions. Future crime DSCM studies where sampling of alternatives would be needed to overcome computational limits could benefit from conducting initial tests with a small subset of choices comparing activity node based importance sampling and simple random sampling to the full model, before opting for one or other sampling strategy. We also encourage further research exploring the circumstances in which importance sampling outperforms simple random sampling to guide crime location choice studies. For example, such research might systematically vary the study area size, number and size of the spatial units, crime type, types of variables (activity node and opportunity related) and data sources. The present results suggest that importance sampling may be particularly important when estimating variables that are idiosyncratic (i.e., that vary simultaneously across alternatives and across offenders, such as awareness space) or have skewed distributions, such as (again) awareness space but also opportunity variables that are highly skewed. Future research could also investigate the effect of decreasing the proportion of alternatives that are importance sampled, to establish the point at which the estimates become unreliable, by comparison to including 100% in the first (5 km) stratum as was done here.

Several limitations of the present data source also warrant acknowledgement. First, the results may only generalize to location choices of offenders who have been identified and proceeded against. If the predictor variables impact the likelihood of the offender being caught, data from solved cases may not be representative of crime location choices of all offenders (Bernasco et al. 2013; Ruiter 2017). Selection bias will exist if, for example, offending near home or prior crime locations makes it more likely that the offender

is caught, or more likely that the offense is reported to the police in the first place. It was not possible with the present dataset to test for these two types of selection bias. However, prior research has found a lack of association between spatial variables and clearance rates (Bernasco et al. 2013; Lammers 2014; Chiu and Leclerc 2020) and that similar sources of bias in police data (in particular reporting rates) have less effect on analysis using larger spatial units like the neighborhoods used in this study (Buil-Gil et al. 2021).

Second, the data do not include all activity nodes of all offenders. The extent to which any offender's pre-offense activity locations are recorded depends on the extent of their prior contact with police, so many activity nodes naturally remain unknown to police. However, given that peoples' current activity nodes tend to cluster together (Golledge 1999; Schönfelder and Axhausen 2002), it is highly likely that the recorded activity locations are indicative of other, latent, activity nodes. The fact that the odds of crime location choice remained significant and large (ORs 16.1 to 21.5) even 2-5 km from activity nodes suggests that the data may indeed capture additional nodes or awareness space. It also suggests that distance bands beyond 5 km from activity nodes may explain additional variance in location choice and should thus be included in future research.

Lastly, the present results are confined to the use of conditional logit rather than other discrete choice models. But other models can be more appropriate when modelling crime location choice. For example nested logit models could better account for decisions made at tiers of spatial units such as neighborhoods and specific houses (Vandeviver and Bernasco 2020) and mixed logit models are useful for accounting for variation in preferences between different offenders (Townsley et al. 2016; Frith 2019). Further, both models also relax the assumption of independence of irrelevant alternatives (IIA) which applies to conditional logit. The IIA assumption requires that the probability of a given alternative being chosen be independent of the characteristics of other alternatives (Ben-Akiva and Lerman 1985). In spatial choice scenarios, it is likely that alternatives are not independent; the choice may be influenced by the presence, or characteristics, of nearby alternatives (Bernasco 2010). In choosing where to commit a burglary, for example, an offender may be more likely to choose a neighborhood with attractive burglary targets that is surrounded by other neighborhoods with attractive targets, than a neighborhood that has the same level of attractive targets surrounded by less attractive neighborhoods. However, there is no proof that sampling from alternatives, randomly or otherwise, produces robust estimates for these models (von Haefen and Domanski 2013). Future crime location choice research might therefore explore means of sampling for these models, following recent developments in sampling methods for them in other domains (Guevara and Ben-Akiva 2013a, b; von Haefen and Domanski 2013).

Conclusion

The findings of this paper have important implications for future crime location choice studies, and make a novel contribution to the growing literature on offenders' spatial decision making. We presented a relatively simple and effective method for importance sampling from alternatives which if adopted in future crime DSCM studies could enable the use of larger datasets (e.g., with more variables, wider study areas, or more granular spatial or spatio-temporal units) to yield greater insights into crime location choice. Our results suggest that future DSCM crime location choice studies with such large datasets should sample from alternatives (rather than sampling from offenders/offenses, which reduces statistical power), and should consider conducting initial tests to determine whether simple

random or importance sampling is optimal. Further, this is the first New Zealand based study in the DSCM paradigm, and the first to specifically examine non-residential burglary and sexual offenses. In finding a strong relationship between the locations of offenders' prior criminal and non-criminal activities and their crime locations, the results support the generalizability of Crime Pattern Theory (Brantingham and Brantingham 1991, 1993) and previous DSCM studies across jurisdictions and crime types.

Appendix 1

See Fig. 2
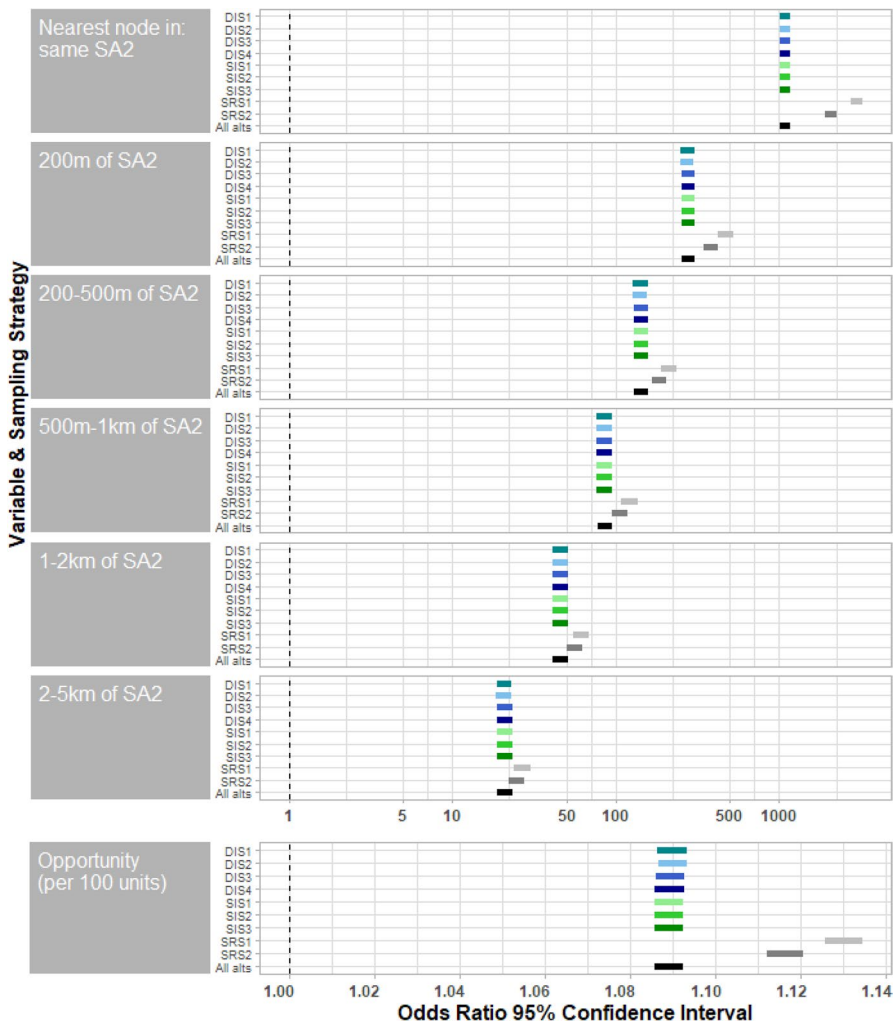
Fig. 2 Parameter estimates for non-residential burglary location choice by sampling strategy

Appendix 2

See Fig. 3

Fig. 3 Parameter estimates for commercial robbery location choice by sampling strategy

Appendix 3

See Fig.

Fig. 4 Parameter estimates for personal robbery location choice by sampling strategy

Appendix 4

See Fig. 5

Fig. 5 Parameter estimates for extra-familial sex offense location choice by sampling strategy
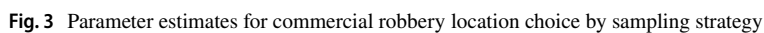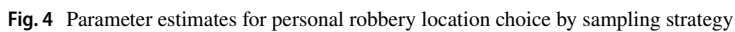
Appendix 5
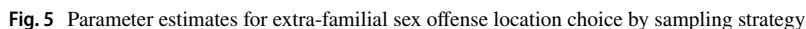
See Fig. 6
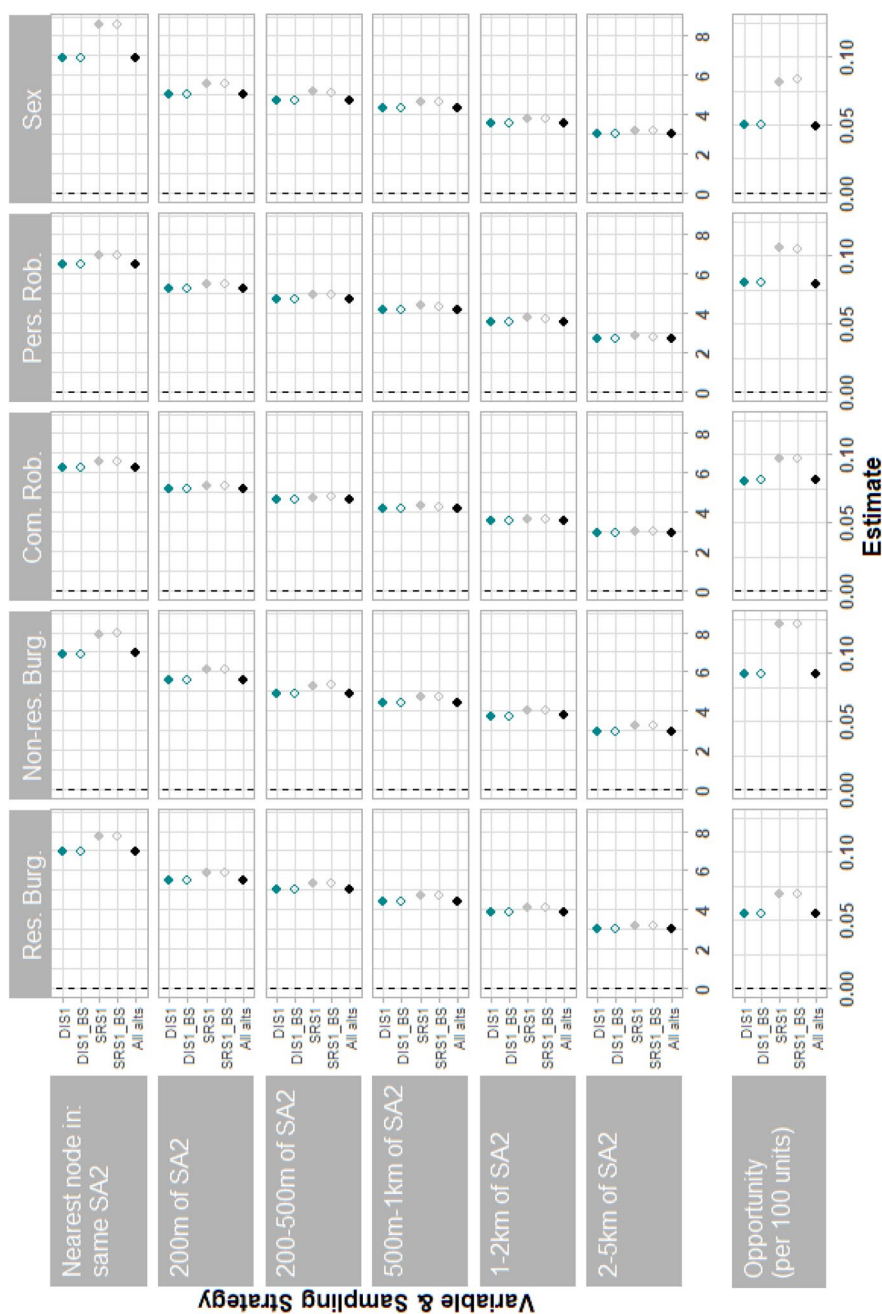
Fig. 6 Parameter estimates for bootstrapped (BS) and non-bootstrapped sampling strategies

Declaration

References

Altizio A, York D (2007) Robbery of convenience stores. U.S. Department of Justice, Office of Community Oriented Policing Services, Washington, DC

Ben-Akiva ME, Bowman JL (1998) Integration of an activity-based model system and a residential location model. Urban Stud 35:1131–1153. https://doi.org/10.1080/0042098984529

Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: Theory and application to travel demand. MIT Press, Cambridge, MA

Bernasco W (2006) Co-offending and the choice of target areas in burglary. J Investig Psych Offender Profil 3:139–155. https://doi.org/10.1002/jip.49

Bernasco W (2010) Modeling micro-level crime location choice: application of the discrete choice framework to crime at places. J Quant Criminol 26:113–138. https://doi.org/10.1007/s10940-009-9086-6

Bernasco W (2017) Modeling offender decision making with secondary data. In: Bernasco W, Van Gelder J-L, Elffers H (eds) The Oxford handbook on offender decision making. Oxford University Press, Oxford, England, pp 569–586

Bernasco W (2019) Adolescent offenders' current whereabouts predict locations of their future crimes. PLoS ONE 14:e0210733. https://doi.org/10.1371/journal.pone.0210733

Bernasco W, Jacques S (2015) Where do dealers solicit customers and sell them drugs? a micro-level multiple method study. J Contemp Crim Justice 31:376–408. https://doi.org/10.1177/1043986215608535

Bernasco W, Nieuwbeerta P (2005) How do residential burglars select target areas? a new approach to the analysis of criminal location choice. Br J Criminol 45:296–315. https://doi.org/10.1093/bjc/azh070

Bernasco W, Block R, Ruiter S (2013) Go where the money is: modeling street robbers' location choices. J Econ Geogr 13:119–143. https://doi.org/10.1093/jeg/lbs005

Bernasco W, Johnson SD, Ruiter S (2015) Learning where to offend: effects of past on future burglary locations. Appl Geogr 60:120–129. https://doi.org/10.1016/j.apgeog.2015.03.014

Bernasco W, Ruiter S, Block R (2017) Do street robbery location choices vary over time of day or day of week? a test in Chicago. J Res Crime Delinq 54:244–275. https://doi.org/10.1177/0022427816680681

Bhat C, Govindarajan A, Pulugurta V (1998) Disaggregate attraction-end choice modeling formulation and empirical analysis. Transp Res Rec 1645:60–68. https://doi.org/10.3141/1645-08

Bichler G, Malm A, Christie-Merrall J (2012) Urban backcloth and regional mobility patterns as indicators of juvenile crime. In: Andresen MA, Kinney JB (eds) Patterns, prevention, and geometry of crime. Routledge, London, England, pp 118–136

Bowman JL, Ben-Akiva ME (2001) Activity-based disaggregate travel demand model system with activity schedules. Transp Res Part A 35:1–28. https://doi.org/10.1016/S0965-8564(99)00043-9

Brantingham PL, Brantingham PJ (1991) Notes on the geometry of crime. In: Brantingham PJ, Brantingham PL (eds) Environmental criminology, 2nd edn. Waveland Press, Prospect Heights, IL, pp 27–54

Brantingham PL, Brantingham PJ (1993) Environment, routine, and situation: Toward a pattern theory of crime. In: Clarke RV, Felson M (eds) Routine activity and rational choice. Transaction Publishers, Piscataway, NJ, pp 259–294

Bright D, Whelan C, Morselli C (2020) Understanding the structure and composition of co-offending networks in Australia. Australian Institute of Criminology Australia

Buil-Gil D, Moretti A, Langton SH (2021) The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis. J Exp Criminol. https://doi.org/10.1007/s11292-021-09457-y

Chiu Y-N, Leclerc B (2020) Predictors and Contexts of Unsolved and Solved Sexual Offenses. Crime Delinq 66:1268–1295. https://doi.org/10.1177/0011128719879027

Clare J, Fernandez J, Morgan F (2009) Formal evaluation of the impact of barriers and connectors on residential burglars' macro-level offending location choices. Aust N Z J Criminol 42:139–158. https://doi.org/10.1375/acri.42.2.139

Curtis-Ham S, Bernasco W, Medvedev ON, Polaschek DLL (2020) A framework for estimating crime location choice based on awareness space. Crime Sci 9:1–14. https://doi.org/10.1186/s40163-020-00132-7

Curtis-Ham S, Bernasco W, Medvedev ON, Polaschek DLL (2021) A national examination of the spatial extent and similarity of offenders' activity spaces using police data. ISPRS Int J Geo-Inf 10(2):47. https://doi.org/10.3390/ijgi10020047

Duncombe W, Robbins M, Wolf DA (2001) Retire to where? a discrete choice model of residential location. Int J Popul Geogr 7:281–293. https://doi.org/10.1002/ijpg.227

Frejinger E, Bierlaire M, Ben-Akiva M (2009) Sampling of alternatives for route choice modeling. Transportation Research Part B: Methodological 43:984–994. https://doi.org/10.1016/j.trb.2009.03.001

Frith MJ (2019) Modelling taste heterogeneity regarding offence location choices. J Choice Modell 33:100187. https://doi.org/10.1016/j.jocm.2019.100187

Frith MJ, Johnson SD, Fry HM (2017) Role of the street network in burglars' spatial decision-making. Criminology 55:344–376. https://doi.org/10.1111/1745-9125.12133

Golledge R (1999) Human wayfinding and cognitive maps. In: Golledge R (ed) Wayfinding behavior: Cognitive mapping and other spatial processes. Johns Hopkins University Press, Baltimore, MD, pp 5–45

Guevara CA, Ben-Akiva ME (2013a) Sampling of alternatives in multivariate extreme value (MEV) models. Transportation Research Part B: Methodological 48:31–52. https://doi.org/10.1016/j.trb.2012.11.001

Guevara CA, Ben-Akiva ME (2013b) Sampling of alternatives in logit mixture models. Transportation Research Part B: Methodological 58:185–198. https://doi.org/10.1016/j.trb.2013.08.011

Guevara CA, Chorus CG, Ben-Akiva ME (2016) Sampling of alternatives in random regret minimization models. Transp Sci 50:306–321. https://doi.org/10.1287/trsc.2014.0573

Hanayama A, Haginoya S, Kuraishi H, Kobayashi M (2018) The usefulness of past crime data as an attractiveness index for residential burglars. J Investigative Psychology and Offender Profiling 15:257–270. https://doi.org/10.1002/jip.1507

Hassan MN, Rashidi TH, Nassir N (2019) Consideration of different travel strategies and choice set sizes in transit path choice modelling. Transportation (dordrecht). https://doi.org/10.1007/s11116-019-10075-x

Huybers T (2005) Destination choice modelling: what's in a name? Tour Econ 11:329–350. https://doi.org/10.5367/000000005774352999

Jonnalagadda N, Freedman J, Davidson WA, Hunt JD (2001) Development of microsimulation activity-based model for San Francisco: destination and mode choice models. Transp Res Rec 1777:25–35. https://doi.org/10.3141/1777-03

Kim J, Lee S (2017) Comparative analysis of traveler destination choice models by method of sampling alternatives. Transp Plan Technol 40:465–478. https://doi.org/10.1080/03081060.2017.1300242

King G, Honaker J, Joseph A, Scheve K (2000) Analyzing incomplete political science data: an alternative algorithm for multiple imputation. American Political Science Review 95:49–69

Lammers M (2014) Are arrested and non-arrested serial offenders different? a test of spatial offending patterns using DNA found at crime scenes. J Res Crime Delinq 51:143–167. https://doi.org/10.1177/0022427813504097

Lammers M (2018) Co-offenders' crime location choice: do co-offending groups commit crimes in their shared awareness space? Br J Criminol 58:1193–1211. https://doi.org/10.1093/bjc/azx069

Lammers M, Menting B, Ruiter S, Bernasco W (2015) Biting once, twice: the influence of prior on subsequent crime location choice. Criminology 53:309–329. https://doi.org/10.1111/1745-9125.12071

Lemp JD, Kockelman KM (2012) Strategic sampling for large choice sets in estimation and application. Transp Res Part A 46:602–613. https://doi.org/10.1016/j.tra.2011.11.004

Li M-T, Chow L-F, Zhao F, Li S-C (2005) Geographically stratified importance sampling for the calibration of aggregated destination choice models for trip distribution. Transp Res Rec 1935:85–92. https://doi.org/10.3141/1935-10

Long D, Liu L, Feng J, Zhou S (2018) Assessing the influence of prior on subsequent street robbery location choices: A case study in ZG city. China Sustain 10:1818. https://doi.org/10.3390/su10061818

McFadden D (1977) Modelling the choice of residential location. Yale University, Cowles Foundation for Research in Economics

McFadden D (1984) Econometric analysis of qualitative response models. In: Griliches P, Intriligator MD (eds) Handbook of econometrics. Elsevier, Amsterdam, The Netherlands, pp 105–142

Menting B (2018) Awareness x opportunity: testing interactions between activity nodes and criminal opportunity in predicting crime location choice. Br J Criminol 58:1171–1192. https://doi.org/10.1093/bjc/azx049

Menting B, Lammers M, Ruiter S, Bernasco W (2016) Family matters: effects of family members' residential areas on crime location choice. Criminology 54:413–433. https://doi.org/10.1111/1745-9125.12109

Menting B, Lammers M, Ruiter S, Bernasco W (2020) The influence of activity space and visiting frequency on crime location choice: findings from an online self-report survey. Br J Criminol 60:303–322. https://doi.org/10.1093/bjc/azz044

Mersman O (2019) microbenchmark: Accurate timing functions. Version 1.4–7URL https://CRAN.R-project.org/package=microbenchmark

Nerella S, Bhat CR (2004) Numerical analysis of effect of sampling of alternatives in discrete choice models. Transp Res Rec 1894:11–19. https://doi.org/10.3141/1894-02

Nevo A (2001) Measuring market power in the ready-to-eat cereal industry. Econometrica 69:307–342

Nguyen HTA, Chikaraishi M, Fujiwara A, Zhang J (2017) Mediation effects of income on travel mode choice: Analysis of short-distance trips based on path analysis with multiple discrete outcomes. Transp Res Rec 2664:23–30. https://doi.org/10.3141/2664-03

Park H, Park D, Kim C et al (2013) A comparative study on sampling strategies for truck destination choice model: case of Seoul metropolitan area. Can J Civ Eng 40:19–26. https://doi.org/10.1139/cjce-2012-0433

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rossmo DK (2000) Geographic profiling. CRC Press, Boca Raton, FL

Rubin D (1987) Multiple Imputation for Nonresponse in Surveys, 1st edn. Wiley, NY

Ruiter S (2017) Crime location choice. In: Bernasco W, Van Gelder J-L, Elffers H (eds) The Oxford handbook of offender decision making. Oxford University Press, Oxford, pp 398–420

Schönfelder S, Axhausen KW (2002) Measuring the size and structure of human activity spaces: The longitudinal perspective. ETH, Zurich

Shiftan Y (1998) Practical approach to model trip chaining. Transp Res Rec 1645:17–23. https://doi.org/10.3141/1645-03

Song G, Bernasco W, Liu L et al (2019) Crime feeds on legal activities: Daily mobility flows help to explain thieves' target location choices. J Quant Criminol. https://doi.org/10.1007/s10940-019-09406-z

Taylor N (2002) Robbery against service stations and pharmacies: recent trends. Australian Institute of Criminology, Canberra, Australia

Therneau T (2020) A Package for Survival Analysis in R. Version 3.1–12URL https://CRAN.R-project.org/package=survival

Townsley M (2016) Offender mobility. In: Wortley R, Townsley M (eds) Environmental criminology and crime analysis. Routledge, London, England, pp 142–161

Townsley M, Birks D, Bernasco W et al (2015) Burglar target selection: a cross-national comparison. J Res Crime Delinq 52:3–31. https://doi.org/10.1177/0022427814541447

Townsley M, Birks D, Ruiter S et al (2016) Target selection models with preference variation between offenders. J Quant Criminol 32:283–304. https://doi.org/10.1007/s10940-015-9264-7

van Daele S, Vander Beken T (2010) Journey to crime of "itinerant crime groups." Policing Int J. 33:339–353. https://doi.org/10.1108/13639511011044920

van Daele S, Vander Beken T, Bruinsma GJN (2012) Does the mobility of foreign offenders fit the general pattern of mobility? Eur J Criminol 9:290–308. https://doi.org/10.1177/1477370812440065

van Sleeuwen SEM, Ruiter S, Menting B (2018) A time for a crime: temporal aspects of repeat offenders' crime location choices. J Res Crime Delinq 55:538–568. https://doi.org/10.1177/0022427818766395

Vandeviver C, Bernasco W (2020) "Location, location, location": effects of neighborhood and house attributes on burglars' target selection. J Quant Criminol 36:779–821. https://doi.org/10.1007/s10940-019-09431-y

Vandeviver C, Neutens T, van Daele S et al (2015) A discrete spatial choice model of burglary target selection at the house-level. Appl Geogr 64:24–34. https://doi.org/10.1016/j.apgeog.2015.08.004

von Haefen RH, Domanski A (2013) Estimating mixed logit models with large choice sets. In: International Choice Modelling Conference. Sydney

Authors and Affiliations

Sophie Curtis-Ham[1] · **Wim Bernasco**[2,3] · **Oleg N. Medvedev**[1] · **Devon L. L. Polaschek**[1]

[1] Te Puna Haumaru NZ Institute of Security and Crime Science & Te Kura Whatu Oho Mauri School of Psychology, Te Whare Wānanga o Waikato University of Waikato, Hamilton 3240, New Zealand

[2] Netherlands Institute for the Study of Crime and Law Enforcement (NSCR), 1081 HV Amsterdam, The Netherlands

[3] Department of Spatial Economics, School of Business and Economics, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands