A discrete choice and configurational analysis of burglary offence location choices

Michael James Frith
University College London

Thesis submitted for the degree of Doctor of Philosophy
Declaration

I, Michael James Frith, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
Abstract

The purpose of this thesis is to explore the applications of configurational methods from the fields of graph theory and space syntax and discrete choice methods from economics to the analysis of crime. In the work that follows, this thesis will argue that, based on current environmental criminology theory, the movement of offenders and ordinary citizens play a vital but under-researched role in the distributions of crime. For offenders, it shapes their awareness and familiarity of the opportunities for crime. For ordinary citizens, it determines the supply of potential bystanders and the quality of ambient guardianship. This thesis will contend that the current methods for empirically describing or estimating both types of movement and the approaches typically used for analysing (their role in) crime patterns are not without significant shortcomings. As such, a series of novel graph theory network measures and a sample of discrete choice methods (the conditional logit, mixed logit and latent class logit models) are explored in relation to these issues. These methods are then jointly employed and empirically tested and compared in a set of original analyses of the burglary location choices in Buckinghamshire (UK).
**Impact statement**

In analysing the novel ways configurational and choice methods can be used and combined to analyse offence location choices, the work contained in this thesis contributes to criminological literature. Particularly, in terms of creating novel movement measures for examining long-standing assumptions (e.g. regarding the effects of familiarity of offence location choices) and debates (e.g. regarding the role of passers-by in crime patterns) within the literature. The value and impact of this work criminology is evidenced by the publication of the paper, “role of the street network in burglars’ spatial decision-making”, in the leading criminology journal: Criminology (Frith, Johnson and Fry, 2017).

Outside of research, undertaking this work has brought about several opportunities to lecture to postgraduate students. In particular, in one module where part of the focus of a computer workshop is on the types of analyses conducted and presented in this thesis. Other aspects of this work have also been presented in other modules.

Outside of academia, the work undertaken as part of this thesis has also contributed to the government and general public in Buckinghamshire as the work was also presented to and discussed with the County Council. Some of the analyses presented in this thesis, and other analyses conducted on behalf of Buckinghamshire County Council were also incorporated in their Joint Strategic Assessment created in 2014. Going forward, the contacts forged through this work also means the (potential) practical impact of this thesis is likely to grow.
Acknowlegements

I would like to thank my two supervisors, Professor Shane Johnson and Dr Hannah Fry, for all of their unlimited guidance and support.

I am also extremely grateful for the practical support and assistance from Buckinghamshire County Council, in particular Cath Marriott, Thames Valley Police, and the financial support from Engineering and Physical Sciences Research Council (EPSRC) under grant no. EP/G037264/1.

Last, and certainly not least, I would like to thank all of my family and friends who have supported me throughout this research (and before!).

Thank you!
Contents

Chapter 1  Introduction ................................................................................. 18
  1.1. Background .................................................................................. 18
  1.2. Outline of the thesis ..................................................................... 19
  1.3. Structure of the thesis ................................................................... 20
  1.4. Associated publications ................................................................. 24

Chapter 2  Criminological Literature Review ............................................ 25
  2.1. Introduction .................................................................................. 25
  2.2. Environmental criminology ........................................................... 28
    2.2.1. Rational choice perspective .................................................... 28
    2.2.2. Routine activities perspective ................................................. 30
    2.2.3. Crime pattern (geometric) perspective ................................... 34
    2.2.4. Social disorganisation and collective efficacy perspectives ...... 37
    2.2.5. Eyes on the street perspective ............................................... 40
    2.2.6. Defensible space perspective ............................................... 40
  2.3. Theoretical model of burglary ....................................................... 42
    2.3.1. Offenders and the road network ............................................. 42
    2.3.2. Passer-by movement ............................................................ 44
    2.3.3. Other factors ........................................................................ 45

Configurational Methods

Chapter 3  Introduction to configurational methods .................................. 48
  3.1. Introduction .................................................................................. 48
  3.2. Perception methods ..................................................................... 49
  3.3. Land use methods ....................................................................... 50
  3.4. Road type methods ..................................................................... 53
  3.5. Configurational methods ............................................................... 55
    3.5.1. Representations .................................................................... 56
    3.5.2. Measures ............................................................................. 61
    3.5.3. Impedances ......................................................................... 64
Chapter 4  Meta-analysis of movement flow estimations using configurational methods ........................................ 81
4.1. Introduction .......................................................... 82
4.2. Methodology .......................................................... 84
    4.2.1. Configurational analyses ....................................... 84
    4.2.2. Observation data ............................................... 86
    4.2.3. Correlation methodology ...................................... 88
    4.2.4. Meta-analysis methodology ................................... 91
4.3. Results ............................................................... 95
    4.3.1. Pedestrian results .............................................. 95
    4.3.2. Vehicular results .............................................. 99
4.4. Discussion .......................................................... 103
Appendix A. List of computed configurational analyses ................. 109

Chapter 5  New configurational measures of movement ............... 111
5.1. Introduction .......................................................... 111
5.2. Modelling non-uniform distributions of trip generators .......... 112
    5.2.1. Non-uniform distributions of trip generators .............. 112
    5.2.2. Vertex-weighted betweenness ................................ 113
    5.2.3. Example calculation ........................................... 114
    5.2.4. Guardians and guardianship ................................... 117
    5.2.5. Local and non-local betweenness ......................... 117
    5.2.6. Example calculation ........................................... 118
5.3. Modelling offender awareness spaces ............................. 121
    5.3.1. Offender awareness spaces ................................... 121
    5.3.2. Idiosyncratic betweenness .................................... 122
    5.3.3. Example calculation ........................................... 124
5.4. Discussion .......................................................... 125

Discrete Choice Methods
Chapter 6  Introduction to discrete choice methods ........................................ 130
  6.1.  Introduction ........................................................................................... 130
  6.2.  Target-based approach .......................................................................... 131
  6.3.  Offender-based approach ....................................................................... 131
  6.4.  Mobility-based approach ........................................................................ 132
  6.5.  Discrete choice approach ....................................................................... 133
      6.5.1.  Conditional logit .................................................................................. 134
      6.5.2.  Mixed logit ........................................................................................ 136
      6.5.3.  Latent class logit .............................................................................. 137

Chapter 7  Comparison of the approaches to analysing offender location
preferences ........................................................................................................ 139
  7.1.  Introduction ............................................................................................ 139
  7.2.  Comparisons of the four approaches to analysis ................................. 140
  7.3.  Methodology .......................................................................................... 147
      7.3.1.  Simulation process ............................................................................. 147
      7.3.2.  Data .................................................................................................. 150
      7.3.3.  Analytic methodology ..................................................................... 153
  7.4.  Results .................................................................................................... 156
      7.4.1.  Target-based analyses ..................................................................... 156
      7.4.2.  Offender-based analyses .................................................................. 160
      7.4.3.  Mobility-based analyses .................................................................. 166
  7.5.  Discussion ............................................................................................... 168

Appendix A.  Multi-level mixed-effects Bernoulli regression of the clearance rates
of crime ............................................................................................................. 170

Chapter 8  Dirty data in offence location preference analyses ..................... 174
  8.1.  Data attrition .......................................................................................... 175
      8.1.1.  Methodology ..................................................................................... 178
      8.1.2.  Results ............................................................................................. 180
  8.2.  Methods for improving target-based and choice-based estimates ........ 183
      8.2.1.  Methodology ..................................................................................... 188
      8.2.2.  Results ............................................................................................. 190
  8.3.  Discussion ............................................................................................... 193
Chapter 9
Systematic review and meta-analysis of offence location choice research ................................................................. 200
9.1. Introduction........................................................................................................................................................................... 200
9.2. Systematic literature review .................................................................................................................................................. 203
  9.2.1. Search strategy ................................................................................................................................................................. 203
  9.2.2. Inclusion criteria ............................................................................................................................................................... 204
  9.2.3. Search results (studies) ...................................................................................................................................................... 204
  9.2.4. Search results (analyses) .................................................................................................................................................. 206
9.3. Meta-analysis and meta-regression methodology ................................................................................................................. 216
  9.3.1. Different effect indices ....................................................................................................................................................... 218
  9.3.2. Different predictor variables .............................................................................................................................................. 219
  9.3.3. Different measures of the predictor variables .................................................................................................................. 221
  9.3.4. Non-independent datasets .............................................................................................................................................. 223
  9.3.5. Other analysis differences .............................................................................................................................................. 225
  9.3.6. Other issues ....................................................................................................................................................................... 227
  9.3.7. Model specifications ......................................................................................................................................................... 228
9.4. Meta-analysis results .............................................................................................................................................................. 229
9.5. Meta-regression results ............................................................................................................................................................ 234
9.6. Discussion ............................................................................................................................................................................... 243

Burglary Offence Location Choice Analyses

Chapter 10
Role of guardianship and familiarity in burglary location choices ............................................................................................... 248
10.1. Introduction............................................................................................................................................................................ 248
10.2. Hypotheses............................................................................................................................................................................ 249
10.3. Methodology ......................................................................................................................................................................... 252
  10.3.1. Study Area ...................................................................................................................................................................... 252
  10.3.2. Choice Set .................................................................................................................................................................... 253
  10.3.3. Crime Data .................................................................................................................................................................. 254
  10.3.4. Graph theory-derived choice criterion ......................................................................................................................... 255
  10.3.5. Other choice criterion .................................................................................................................................................. 258
  10.3.6. Choice Model .............................................................................................................................................................. 262
10.4. Results.................................................................................................................................................................................. 264
10.4.1. Guardianship ............................................................... 265
10.4.2. Familiarity ............................................................... 266
10.5. Discussion ........................................................................ 266

Chapter 11  Preference variation in burglary location choices .......... 269
11.1. Introduction ....................................................................... 269
11.2. Methodology ..................................................................... 271
  11.2.1. Data ............................................................................. 271
  11.2.2. Choice Model Estimations ............................................. 271
11.3. Results .............................................................................. 272
  11.3.1. Model comparisons ....................................................... 275
  11.3.2. CL Results .................................................................... 277
  11.3.3. ML Results .................................................................... 277
  11.3.4. LC Results .................................................................... 278
11.4. Discussion .......................................................................... 279

Chapter 12  Discussion .................................................................. 282
12.1. Summary ............................................................................ 282
12.2. Key contributions .............................................................. 285
  12.2.1. Configurational and discrete choice methods ............... 285
  12.2.2. Role of movement in offending ..................................... 286
  12.2.3. Burglary location choices ............................................ 287
12.3. Key limitations and avenues for future work ....................... 288
  12.3.1. Configurational methods ............................................ 288
  12.3.2. Offence location choice analyses ............................... 289
  12.3.3. Practical applications of the findings ......................... 290

References ............................................................................. 291
List of Tables

Table 3.1: The observed and perceived changes in vehicular traffic between 1976 and 1977. Adapted from Fowler Jr (1979). .......................................................... 50

Table 3.2: The four best-fitting land-use models (and the variables used within each model) for predicting pedestrian and vehicular traffic........................................ 52

Table 3.3: The observed mean hourly numbers of vehicles and pedestrians for different classes of roads.................................................................................................................. 55

Table 4.1: Summaries of the key parameters in configurational analyses .................. 83

Table 7.1: The estimated odds-ratios of where burglars live in Buckinghamshire and The Hague (Bruinsma et al., 2013)................................................................. 142

Table 7.2: Summary information of the independent variables used in the analyses....153

Table 7.3: List of police.uk crime outcome categories and their classification as ‘cleared’, ‘not cleared’ or ‘N/A’ ........................................................................................................ 171

Table 8.1: The estimated odds-ratios of where burglaries are cleared in Buckinghamshire ...................................................................................................................... 179

Table 9.1: Summaries of the 26 studies and their containing analyses as identified through the systematic literature review......................................................... 208

Table 9.2: Summary information of the unique (144) analyses and studies found through the literature review............................................................................................... 211

Table 9.3: Vote count summary of ten of the most frequently used variables in the offence location choice literature.................................................................................. 215

Table 9.4: Results from Wald tests comparing effect estimates from the meta-analysis ............................................................................................................................. 232

Table 9.5: Results from Wald tests comparing effect estimates from the meta-regression ......................................................................................................................... 238
Table 10.1: Summaries of the pedestrian/choice and vehicular networks ..................256
Table 10.2: Summary statistics of the independent variables (at street segment level) ..258
Table 10.3: Results from the mixed logit analysis of burglary offence location choices  .................................................................264
Table 11.1: Model fits for LC models with two to six classes .................................273
Table 11.2: Results from the conditional logit, mixed logit, and latent class analyses of burglary offence location choices..................................................274
List of Figures

Figure 2.1: The event model of a residential burglary in a middle-class area. Adapted from Clarke and Cornish (1985). .................................................................30

Figure 2.2: The crime triangle showing the convergence of an offender and a target in the absence of guardianship necessary for a crime to occur. Adapted from Felson (1998)........................................................................................................................................31

Figure 2.3: The problem analysis triangle showing the relationships between targets, offenders and places, their controllers and their super controllers. Adapted from Sampson et al. (2010). ..................................................................................................................................................................................33

Figure 2.4: The likely distribution of crime due to the convergence of awareness space and areas with many potential targets. Adapted from Brantingham and Brantingham (1981b). ..................................................................................................................................................................................36

Figure 2.5: The concentric zone model of urban socio-structures. Adapted from Park et al. (1925/1967). ..................................................................................................................................................................................38

Figure 2.6: The defensible space model and the demarcation and hierarchy of public to private. Adapted from Newman (1972). .................................................................................................................................42

Figure 3.1: Maps illustrating the construction of primary and dual graphs (using street segments) of the street network of a village (near Chesterfield, UK). .................57

Figure 3.2: The street network of a village (from Figure 3.1) represented with different units. ...........................................................................................................59

Figure 3.3: The street network of a town (Chesterfield, UK) analysed with six different measures. ........................................................................................................62

Figure 3.4: Shortest paths between two locations according to different impedances and with the street and path network constructed from different units. ..........66

Figure 4.1: Maps of the street segment (left) and axial line (right) road and path networks for the study area including an 8km buffer (see also top inset) and an extract from the Isle of Dogs (bottom). ................................................................................................................................................................85
Figure 4.2: Maps of all pedestrian observation locations by dataset/geographic study-area (left) and for each set of study sub-areas (right) .................................................................90

Figure 4.3: Maps of the vehicular observation locations by the original dataset (left) and by the categorised geographic study-area (right) and study sub-area (right inset). 91

Figure 4.4: Forest plot showing the pooled correlation estimates of the overall and by parameter correlations with pedestrian (top) and vehicular (bottom) traffic ........96

Figure 4.5: Forest plots showing the modulus of the pooled correlation strength estimates between selected specific analyses (permutations of parameters) and pedestrian traffic .................................................................................................................................99

Figure 4.6: Forest plots showing the modulus of the pooled correlation strength estimates between selected specific analyses (permutations of parameters) and vehicular traffic .................................................................................................................................102

Figure 5.1: Two example networks (a and c) and the associated betweenness values when using the original betweenness measure (b and d) and the adjusted betweenness measure (b and e) .................................................................................................................................116

Figure 5.2: An example network (a) and the associated (overall) betweenness values (b), local betweenness (c), neither local or non-local (d) and non-local (e) values (calculated from x₁) .................................................................................................................................120

Figure 5.3: An example network (a), the shortest routes to all other vertices from x₁ (b) and x₄ (d) and the corresponding idiosyncratic betweenness values (c and e) ....126

Figure 7.1: Examples of the simulated offender and offence datasets ..............................................151

Figure 7.2: The percentage of locational characteristic variables falsely detected as significant predictors of offending by locational characteristic, dataset size and effect size (of proximity) in target-based and choice-based analyses (hypothesis 1) .................................................................................................................................157

Figure 7.3: The percentage of locational characteristic variables undetected as significant positive predictors of offending by locational characteristic, dataset size and effect size (of all variables) in target-based and choice-based analyses (hypothesis 2) ...159
Figure 7.4: The percentage of offences across the first 10 distance bands in two example datasets (top) and the average across all datasets (bottom) using offender-based and choice-based analyses (hypotheses 3 and 4) .......................................................... 161

Figure 7.5: The percentage of errors found, in terms of negative relationships between proximity and offending (hypothesis 3) and the presence of buffer spaces (hypothesis 4), in the distance decay patterns from offender-based and choice-based analyses ................................................................. 162

Figure 7.6: The percentage of offences across the first 10 distance bands in two example datasets (top) and the average across all datasets (bottom) using offender-based and choice-based analyses (hypotheses 5 and 6) ................................................................. 163

Figure 7.7: The percentage of errors found, in terms of significant relationships between proximity and offending (hypothesis 5) and the presence of buffer spaces (hypothesis 6), in the distance decay patterns from offender-based and choice-based analyses ........................................................................ 164

Figure 7.8: The percentage of offences across the first 10 distance bands in two example datasets (top) and the average across all datasets (bottom) using offender-based and choice-based analyses (hypotheses 7 and 8) ................................................................. 165

Figure 7.9: The percentage of errors found, in terms of significant differences between the distributions of offences and that expected by the effect size of proximity (hypothesis 7) and the presence of buffer spaces (hypothesis 8), in the distance decay patterns from offender-based and choice-based analyses ......................................... 166

Figure 7.10: The percentage of variables that go undetected as significant positive/negative predictors of offending by the variable, dataset size and effect size in mobility-based and choice-based analyses (hypothesis 9) ........................................ 167

Figure 7.11: The percentage of offences by distance band for arsons (left) from Rossmo (2000) and residential burglaries (right) from Canter and Youngs (2008) ............. 169

Figure 8.1: Sankey chart of the attrition of 100 burglaries in the UK criminal justice system ............................................................................................................................. 176
Figure 8.2: The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to target-based analyses and choice-based analyses using randomly sampled and non-randomly sampled *cleared* burglaries (hypotheses 1-5). .......................................................... 181

Figure 8.3: The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to three versions of target-based analyses (hypotheses 6 and 7). ......................... 190

Figure 8.4: The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to three versions of choice-based analyses (hypotheses 8 and 9). ............................... 192

Figure 8.5: The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to the best performing target-based and choice-based analyses (hypothesis 10)...... 193

Figure 8.6: The percentage of variables that go undetected as significant positive predictors of offending when using different clearance rates of datasets containing 1,000 reported and recorded burglaries according to the tested target-based and choice-based analyses................................................................. 194

Figure 8.7: Contour charts showing the percentage of variables that go undetected as significant positive predictors of offending for datasets generated with different sample and effect sizes according to the tested target-based and choice-based analyses................................................................................................................. 196

Figure 8.8: A decision tree for selecting the approach and model for analysing offence location choices................................................................. 197

Figure 9.1: Sankey diagram showing the breakdown of studies identified and excluded based on each of the criteria................................................................. 205

Figure 9.2: Euler grid diagram showing the studies found through each element of the search strategy .................................................................................. 206

Figure 9.3: Diagram of the spatial units – in terms of size and the number of residents – used in previous discrete choice analyses of offence location choices.......... 213
Figure 9.4: Forest plot of the findings from meta-analysis results

Figure 9.5: Forest plot of the findings from meta-regression (black) and meta-analysis (grey)

Figure 9.6: Euler grid diagram showing the meta-regression results including the average effects of each variable and the effects of each moderator variable on the effects of each variable

Figure 10.1: The expected effects on crime of local, neither local or non-local, and non-local movement according to three theories

Figure 10.2: The expected effects on crime of local, neither local or non-local, and non-local movement according to three theories

Figure 10.3: Map of the county of Buckinghamshire study area

Figure 10.4: Estimates of socioeconomic heterogeneity at the OA-level (a), as an interpolated surface (b) and then assigned to street segments (c) in Buckinghamshire. Inset maps are of Aylesbury (top right) and High Wycombe (bottom left).

Figure 10.5: The expected and observed effects on burglary location choices of local, neither local or non-local, and non-local movement according to three theories

Figure 11.1: Estimated ML and LC distributions of the coefficients
Chapter 1

Introduction

1.1. Background

This thesis concerns the geographic distribution of burglary offence locations and modelling the underlying spatial decision-making process. In particular, it will explore how the movements of the offender and ordinary citizens contribute to burglary offence location choices. In terms of methodological contributions, it will do this by exploring and developing the criminological applications of configurational methods from the fields of graph theory and space syntax, and discrete choice methods from economics.

In the work that follows, this thesis will argue that, based on current environmental criminology theory, the movement of offenders and ordinary citizens play a vital but under-researched role in explaining the distribution of crime. For offenders, it shapes their awareness and familiarity of crime opportunities. For ordinary citizens, it determines the supply of potential bystanders and the quality of ambient guardianship. I will argue that existing methods for empirically describing or estimating both types of movement, and the approaches typically used for analysing (their role in) crime patterns, are not without significant shortcomings. As such, a series of novel street network metrics and a sample of discrete choice methods – used to examine offender decision making - are
explored in relation to these issues. These methods are then jointly employed and empirically tested and compared in a set of original analyses of burglar location choices in Buckinghamshire (UK).

The principal aims and contributions of this work are threefold. The first, and primary technical contribution, is to demonstrate the utility of these methods, including their benefits over alternative and more established methods. The second is to provide greater insight into the role of offender and ordinary citizen movement in crime location choices than has been possible hitherto. The third is – after accounting for these influences - to generate more accurate estimates of the influence of other factors that affect crime pattern formation. These constitute the key technical and theoretical contributions of this work.

1.2. Outline of the thesis

The remainder of this thesis is organised into 11 chapters. Chapter 2 provides a review of the criminological literature that motivates the remainder of this thesis. Chapters 3-11 belong to three general themes, and are consequently organised divided into three sections:

- Chapters 3-5 constitute the Configurational Methods section and are all related to the introduction and application of “configurational methods” to estimating movement along street networks;

- Chapters 6-9 represent the Discrete Choice Methods section and are all related to the introduction and application of discrete choice methods to analysing offence location choices; and

- Chapters 10 and 11 compose the Main Analysis section and concern the simultaneous use of both methods in the analyses of burglary location choices in Buckinghamshire (UK).
For the first two of these sections, most of the relevant material concerning the background to the methods, their theoretical underpinnings, technical explanations, and prior literature are reviewed in the first chapters of each section (Chapters 3 and 6). The exception to this is Chapter 5 in which novel (configurational) methods are proposed. In this case, the material solely related to those methods are discussed in that chapter. Chapter 12 provides a Summary and Discussion of the material covered in the thesis.

1.3. Structure of the thesis

In Chapter 2, a review of the relevant criminological literature is provided which begins with a general background to the study of crime. This chapter will argue that “traditional” criminological theory is poorly suited to understanding the spatial distributions of crime. The rest of this Chapter will then give an overview of environmental criminology and how its focus on the crime setting provides a suitable framework for understanding where crime events occur. The key theories and perspectives are briefly described and then summarised together, along with the results of earlier research. In doing so, a set of testable hypotheses regarding burglar location choices are articulated. In particular, the role of people’s movement patterns is discussed as an explanation for crime pattern formation. The key disagreements between theories and the paucity of research about the role of movement in crime locations are also discussed.

Chapter 3 introduces the key methods from the criminological and wider literature for counting or estimating (aggregate) movement flows. The limitations of each approach are identified and discussed, and it is argued that configurational methods offer the most appropriate solution for estimating movement patterns in statistical models of crime. These configurational methods, or more specifically graph theoretical and space syntactic configurational methods, are then discussed in more detail. The relevant terminology and technical material are introduced.
The configurational methods presented in Chapter 3 can be computed and the road networks analysed in many ways. Chapter 4 is concerned with this and assessing the accuracy of each approach in predicting movement flows. These findings inform the choice of method to be used in the rest of this thesis and that which is recommended for future research. To do this, a systematic review of this literature is first conducted and briefly described. The findings of this review highlight serious shortcomings in the existing literature and suggest the need for more useful estimates of the accuracy of the approaches considered. As such, using movement flow data from London (UK), a set of original correlations between the different types of analyses and pedestrian and vehicular movement flows are computed. The results are then synthesised in a meta-analysis and the implications of the findings are discussed.

While the types of analyses, and specifically a measure of “betweenness”, found in Chapter 4 can be used to estimate aggregate movement flows, Chapter 5 proposes three extensions. The first proposes an approach to improve the accuracy of the estimates of movement flows by employing a weighting strategy that takes account of the fact that movement is more or less common from and to certain locations around the network. Drawing from various theoretical perspectives, two further extensions are proposed to provide nuanced estimates of movement flows intended to provide insight into the “guardianship” potential that those moving through the network might provide, and the parts of the street network where offenders are most likely to be familiar. These extensions are explained, justified, and then illustrated with an example street network.

Chapter 6 introduces the four general statistical approaches: the target-based, offender-based, mobility-based, and choice-based, for analysing offence locations and offender preferences. The limitations of each approach are identified and discussed, and it is argued that the latter approach, the choice-based approach, is theoretically superior to other
approaches and offers for the best solution for analysing crime patterns. The most common discrete choice model, the conditional logit [CL], is then discussed. Given the limitations of the CL, in particular that it assumes all offenders share the same preferences which is counter to much criminological research, two further models that allow for heterogeneity between offenders, the mixed logit and latent class logit, are then explored. The relevant terminology and technical material are also explained.

While the discrete choice approach is theoretically superior to the other three approaches, there has been no (formal) comparisons within criminology of the results of analyses using these different approaches. As such, the impact of the analytical approach on the findings and their interpretation is currently unknown. Exploring this is the focus of Chapters 7 and 8. To facilitate this, synthetic burglary data are generated using theoretical parameters and each of the methods are assessed to examine the extent to which the inferences drawn using the methods match the data generating process employed to produce the data (Chapter 7). Chapter 8 furthers these analyses and introduces the concept of ‘dirty data’ to the data generating process. That is, the fact that real-world crime data are imperfect. For example, because not all crimes are reported to, recorded by, or cleared by the police. As such, and because each approach requires different types of data, the dirtiness of data can impact their findings. Methods for improving the ability of the approaches to detect the parameters used to generate the data are also tested and the results show the general superiority of the choice-based approach.

The findings from Chapters 7 and 8 support the application of the choice-based approach to analysing offence location choices, however, there are many ways in which the approach can be applied. This includes the variables that are used, or the specific choice model applied. Chapter 9 is concerned with the analysis of the effects of these differences. This chapter quantitatively syntheses the results from existing offence location choice
studies to estimate the effects of using, or not using, different parameters, variables, and model types. To this end, a systematic review of the literature is conducted, and the resulting studies are synthesised together using a series of meta-analyses and meta-regressions. Although the results are limited by the types and number of available analyses, the results provide insight into which parameters should be considered and included in future analyses (of offence location choices). This also informs the analyses presented in Chapters 10 and 11 of this thesis.

In Chapters 10 and 11, the configurational and discrete choice methods from Chapters 3-5 and 6-9 respectively are employed together in a series of analyses of burglary location choices, using data for the county of Buckinghamshire (UK). More specifically, Chapter 10 concerns the street network measures devised in Chapters 4-6 and uses a basic discrete choice model (the conditional logit) to test the measures and hypotheses proposed in Chapter 5. The results largely reconcile with the theoretically expected effects of movement, which demonstrates the premises of those measures.

The analyses of the relationship between movement and the other factors on burglary location choices is extended in Chapter 11 to include two more-sophisticated discrete choice models. These are the mixed logit and latent class models that, for one, incorporate the fact that offenders may differ from each other in terms of their offending preferences. The results again show the significant role movement (of the offender and citizens as bystanders) plays in burglar location choices. The results also reveal in more detail than previous studies how burglars (or offenders more generally) can differ from each other in terms of their offending preferences. The latent class model provides new insight into how these preferences can be discontinuous and offenders grouped into distinct types.

Chapter 12, which concludes the thesis, begins by summarising the findings of the earlier chapters. The key themes, across chapters, in this work are then discussed and the
implications and contributions of each are explored. Lastly, the overall scope of this thesis is established and open questions or expected directions for further work are examined.

1.4. Associated publications

The work presented in this thesis draws from or is otherwise published (or awaiting publication) in three publications. More specifically, and in order of their appearance in the thesis, content in Chapter 3 regarding configurational methods draws from the publication of “using qualitative distance metrics in space syntax and configurational analyses” presented at the 11th Space Syntax Symposium (Frith, 2017). Content in Chapter 6 and the analyses in Chapters 10 and 11 regarding discrete choice methods and the mixed logit analyses of preference heterogeneity draws from and extends the analyses in the publication of “role of the street network in burglars’ spatial decision-making” in Criminology (Frith, Johnson and Fry, 2017). Lastly, and similarly but also regarding the latent class (and mixed logit) analyses of preference heterogeneity, content in Chapter 6 and analyses in Chapters 10 and 11 draw from the forthcoming (under revision) publication of “a latent class discrete choice analysis of heterogeneity in offence location preferences” in the Journal of Choice Modelling (Frith, under review).
Criminological Literature Review

This chapter provides a review of the relevant theoretical and empirical literature concerning crime and why it distributes in space and time. The chapter begins with a broad overview of the general study of crime before introducing environmental criminology and the key related theories. These theories are then summarised to give a comprehensive explanation of burglary location choices and to highlight the theoretically key role of movement in explaining the distribution of crime. In relation to this, disagreements between theories and gaps in the literature are identified and hypotheses to be investigated in this thesis described.

2.1. Introduction

The study of crime and criminal behaviour has an extensive history involving a wide range of frameworks and perspectives. Although exceptions exist, for much of this history the field has been dominated by what can be termed traditional criminology. Here, and although the philosophical standpoint can range from classicism where individuals have free will to engage in crime to positivism where individuals’ engagement in crime is governed by factors outside their control, traditional criminology concentrates on criminality and the criminal
Criminological Literature Review

Disposition. That is, it looks to find and explain why individuals engage in crime using a range of factors such as psychological and biological differences (classical criminology) and social conditions and developmental experiences (positivist criminology). This is with the goal of identifying persons at risk of committing crime and preventing or reducing their engagement or re-engagement.

Although not undervaluing the role and contributions of traditional criminology (for example, see Kautt and Pease, 2013), its presence as the principal or lone framework for understanding crime is disputable. For one, and also called the fundamental attribution bias (Ross, 1977; see also Tilley and Sidebottom, 2015), it arguably errs by emphasising dispositional determinants of crime and downplaying the potential of situational causes. In this way, criminality is regarded as being determined by inherent characteristics or traits that distinguish criminals from non-criminals. This is overly simplistic. For example, a 2003 UK survey on self-reported offending (Budd, Sharp and Mayhew, 2005) suggests much of the population commit crime (41%) but do so infrequently (17% only once in their lifetime and 60-65% fewer than four times) and not persistently (5% of over 25-year-olds had offended in the past year compared to 22% of those under 25). Traditional criminology can also be criticised for the practicality and applicability of its findings. That is, and even noted 40 years ago by Wilson (1975), its distal causes of crime, such as deprivation, are unlikely to be completely eradicated; and even if they are, the policy relevance is questionable as the net effects could only be realised in the far future and in the next wave of adolescents and adults.

In response to these limitations and the gaps in the literature, an alternative framework called environmental criminology started to emerge in the 1970s and 1980s. Although the term itself was previously used in a related work by Jeffery (1971); it was specifically introduced in a book of the same name by Paul and Patricia Brantingham (1981a). Its roots, however,
can be traced to several earlier works. Key historical examples include those in the 19th century by Balbi and Guerry (1829), Guerry (1833/2002) and Quetelet (1842) who analysed spatial and temporal crime patterns using large administrative areas (departments) in France, and Glyde (1856) who found spatial heterogeneity in crime within similar administrative divisions (the county of Suffolk) in England (see also Weisburd, Bruinsma and Bernasco, 2009). Relatively more recent work such as that by the Chicago School from the 1920s and architectural criminologists such as Oscar Newman in the 1970s were also major influences (for a review see Bruinsma and Johnson, 2018).

The emphasis of environmental criminology is the crime setting and the where, when and how crime events occur (Brantingham and Brantingham, 1981a). That is, and unlike traditional criminology that concentrates on the offender and distal causes of crime (criminality), environmental criminology focuses on the proximal determinants of crime and the immediate circumstances of the crime (see also Clarke, 2004). It proposes that crime events are the result of person-situation interactions where situational factors and the environment fundamentally shape the distribution of crime. In this way, and although the offender is still an important aspect, the role and importance of their criminal disposition is rendered secondary; it is the (criminogenic) circumstances that lead to its translation into a crime that are of interest. The goal of this framework is, therefore, the manipulation of these proximal factors to inhibit the commissioning of crime. The successes of which have been shown repeatedly; for example, see Cozens et al. (2005) for a review of one application of environmental criminology, crime prevention through environmental design (see also Clarke, 2010; Welsh and Taheri, 2018).

In what follows, the three main theoretical perspectives of environmental criminology: rational choice (Clarke and Cornish, 1985), routine activity (Cohen and Felson, 1979) and crime pattern (geometry of crime) (Brantingham and Brantingham, 1981b) are discussed. Other
important and related perspectives such as *social disorganisation* (Shaw and McKay, 1942) and *collective efficacy* (Sampson, Raudenbush and Earls, 1997) and *eyes on the street* (Jacobs, 1961) and *defensible space* (Newman, 1972) are also described. They are then synthesised together. Lastly, the key testable discordances between perspectives and gaps in the empirical literature, which are to be examined in this thesis, are outlined.

### 2.2. Environmental criminology

The environmental criminology framework encompasses several perspectives that share the common goal of understanding crime events. Unlike perspectives that are found in many other fields, these are largely mutually compatible and, for example, explain crime events at various levels. This includes the macro or society-wide level, the meso or neighbourhood-wide level and the micro or individual level. In many cases, they also build upon each other. These perspectives are now described in turn.

#### 2.2.1. Rational choice perspective

Although first formally presented by Ronald Clarke and Derek Cornish in 1985, the roots of the rational choice perspective of offending can be traced to much earlier work. This includes that by the early utilitarians such as Caesare Beccaria (1764/1775) and Jeremy Bentham (1823); but it was arguably the economist, Gary Becker, who stimulated interest in this perspective. In Becker’s paper (1968), he postulated that the choice to offend is a rational one and not qualitatively different from other non-crime related choices. That is, an individual will choose to commit an offence if the expected utility (or monetary gains in this version) from doing so exceeds what they would receive from alternative actions. From this, a function regarding the frequency of offending can be described using the expected utility formula (see also von-Neumann and Morgenstern, 1947). This includes the probabilities and utilities of successful and unsuccessful (in terms of being apprehended
or apprehended and convicted) offending and of alternative legal and illegal activities; and other variables, for example, which represent their willingness to offend.

Although related in that the offender still evaluates the choice to offend – here, conceptualised and described in terms of the likely rewards, risks, and effort of doing so – Clarke and Cornish’s rational choice perspective departs from the economic model on several grounds. Although some are minor (see, for example, Andresen, 2014 for his responses), others are non-trivial. For one, they argue against Becker’s emphasis on monetary gains and instead suggest offenders can be motivated by other sources of utility such as for the prospect of excitement or fun (and so can be used to explain expressive crimes). Clarke and Cornish also contend the perfectly optimising decision-making process implied in the economic model. Instead, they suggest that offenders act with limited rationality, for example, because the decision to offend is not always (perfectly) planned. In this sense, it follows bounded rationality as described by Simon (e.g. 1957, 1986) who argues the full rationality models (such as above and traditionally used in economics) are unrealistic. Instead, and due to limitations in time for processing, cognitive resources and available information, people (offenders) may use heuristics or opt for a satisficing solution. Furthermore, this bounding of rationality ignores affect which can cause further deviations from the types of decision-making logic expected in the economic model (see Van Gelder et al., 2013).

Another key difference between the two models is their focus. That is, although the Becker economic model (1968) briefly considers individual decisions to offend and a function to relate the total number of offences to a single person, it is largely concerned with the overall supply of offenses to determine optimal law enforcement. As such, it can be described as a macro-level model. On the other hand, Clarke and Cornish (1985) adopt a crime event-specific focus, and while they produce additional models on the
commencement, continuation and desistance of offending, they emphasise the micro-level in explaining the occurrence of individual crime events. For example, as shown in Figure 2.1, for a residential burglary, the offender evaluates and selects the offence neighbourhood and dwelling (a hierarchical decision process; see also Brown and Altman, 1982) based on the likely effort (e.g. easily accessible and patio doors), risks (e.g. few police patrols and no one at home) and rewards (e.g. especially affluent).

2.2.2. Routine activities perspective

The routine activities perspective was first presented by the sociologists Lawrence Cohen and Marcus Felson in their 1979 article, *social change, and crime rate trends: a routine activity approach*. As presented it gives a macro-level explanation of the distributions and levels of crime by considering the broad effects of societal changes. In many ways, it is also applicable as a micro or meso-level model as it lays the groundwork and describes the circumstances by which crime will occur. That is, and drawing on the classic ecological

**Figure 2.1:** The event model of a residential burglary in a middle-class area. Adapted from Clarke and Cornish (1985).
work by Hawley (1950), Cohen and Felson (1979) argue that offences (or crime events) can only transpire if three elements coincide in time and space. As shown in the *crime triangle* in Figure 2.2, these are an inclined and capable *offender*, a suitable *target*, and the absence of *capable guardians*. The latter include the police but also ordinary citizens and objects such as CCTV cameras, who can prevent the offence. They argue that these convergences will naturally occur as all three go about their everyday lives and *daily routines*. That said, and although potential victims’ legal (and non-legal) daily routines, such as work, school, and leisure, will bring them into risky situations, because offenders must pursue victims (rather than vice versa), it is the structure of the former that may most strongly determine the distribution and quantity of crime.

In proposing this perspective, Cohen and Felson (1979) draw upon and cite various previous works. For example, descriptive analyses in Scarr *et al.* (1973) and Reppetto (1974) which note a relationship between residential crime and the likely rhythms and tempos of residents’ daily activities. The authors test the macro-level predictions of

**Figure 2.2: The crime triangle showing the convergence of an offender and a target in the absence of guardianship necessary for a crime to occur.** Adapted from Felson (1998).
routine activities against the increases in crime and economic conditions in post-World War II United States. They find the crime rates may be explained by broad changes in the social structure and the shift in routine activities from around the home to outside it. For example, increases in lone person households and the participation of women in the workforce which leave more households vacant and guardian-less for longer; particularly during the daytime. They also briefly consider the role of the growing market for consumer goods and their miniaturisation that leaves more goods that are suitable and available to steal.

Since this seminal article, several contributions extending it have been put forward. One example which incorporate aspects of the rational choice perspective regards the suitability of targets. In this work, two acronyms have been suggested to assess the suitability of a target (principally for acquisitive crime). The first, as proposed by Felson and Clarke (1998), and based on some of the terms used in the original article (Cohen and Felson, 1979) is VIVA. Here, the value, inertia, visibility, and access of a target influence (and can assess) its risk of criminal attack. Later, another popular acronym, CRAVED, which built upon VIVA using Mike Sutton’s (e.g. 1998) work on the trading of stolen goods, was proposed by Clarke (1999). Standing for concealable, removable, available, valuable, enjoyable, and disposable, the key differences from VIVA is the distinction of enjoyable (in terms of gaining non-monetary pleasure) from valuable; and the emphasis on the ease of hiding an item (concealability) and ease of which it can be sold or enjoyed (disposability).

A second set of developments regard the guardian component and the notion that it is just one type of crime controller (which primarily functions by protecting possible targets). Drawing on control theory (Hirschi, 1969), Felson (1986) adds handlers, such as parents and probation officers, as another type who can regulate crime by exerting influence over
potential offenders to prevent them from offending. Eck (1994) also adds managers, such as premise owners or employees of premises, who may not be explicitly interested in restraining offenders (handlers) or protecting targets (guards), but their presence and interest in the premise will impede crime. Much later, Sampson et al. (2010) also add super controllers, such as regulatory bodies, which function by forcing or incentivising the controllers to prevent crime. This multi-tier network is illustrated in the problem analysis triangle in Figure 2.3.

Separate work has also focused on the guardians themselves including the distinction between formal guardians such as the police who have a legal remit, informal guardians including residents and passers-by who do not, and semi-formal guardians such as neighbourhood watch groups who exist somewhere between. Felson (1995) also highlights the different levels of responsibility guardians may take at different times. These include personal responsibility that may regard a person’s own property or areas that are

Figure 2.3: The problem analysis triangle showing the relationships between targets, offenders and places, their controllers and their super controllers. Adapted from Sampson et al. (2010).
intimately known, *diffuse responsibility* for areas where they may have some *affiliation*, and *general responsibility* for areas where they have no attachment. Going further, Reynald (2011) suggests these can be interpreted as an individual’s likely capability as a guardian is some function of their physical and social proximity.

2.2.3. Crime pattern (geometric) perspective

Although continuing the traditions of the earlier environmental criminology perspectives, the crime pattern perspective of crime specifically takes a spatial-temporal approach to understanding crime events. Proposed by Patricia and Paul Brantingham (1981b, see also 1993b), a central proposition of this perspective concerns the *environmental backcloth* which represents the elements of the physical environment that mediate the distribution of crime. One product of this backcloth is an offender’s *awareness space*. To explain, consider that outside of offending, offenders are just like non-offenders and travel around the city during their daily routines. This will include going to, from and between *activity nodes* such as their home, work and school locations and shopping and other leisure venues. Based on this movement, or *activity space*, an offender develops an idiosyncratic *awareness space* of crime opportunities. This will include areas around their activity nodes and along the routes between them (*paths*) (see also Figure 2.4) but also areas around them as they search for targets. Given this, that it needs resources such as time to travel and less information is likely known about distant or unexplored locations and so may be adjudged as riskier due to the uncertainty, an offender is likely to avoid exploratory offending forays. They will instead likely prefer offending in known and proximate areas within their awareness space. This effect of the awareness space is also expected independent of the fact that this movement will bring them into direct contact with crime opportunities.

Although somewhat implicit in this description, an offender’s awareness space does not need to develop linearly. Elements of the environmental backcloth can significantly
distort or shape it and many examples are given in the original Brantingham and Brantingham paper (1981b). One example is the street layout. For example, consider that grid layouts are more predictable, and their use will often result in shorter paths compared to organic layouts with windy roads and cul-de-sacs. It is likely that people, including offenders, will use grid layouts more often, including as through paths. As such, offender awareness spaces will be biased towards these areas. As an alternative, if fewer people use areas with organic layouts they may become attractive to offenders (as there will be fewer guardians) which will bias their target searches and their awareness space. The form of movement and the nature of any connectors (Clare, Fernandez and Morgan, 2009) will also likely have an effect. For example, for areas primarily served (from or to) by high-speed or underground mass-transit systems the awareness spaces may be nodal with large gaps in awareness between stations. In comparison, for areas more reliant on buses it may be expected that awareness spaces will follow the bus routes but be less nodal with more knowledge along the paths. Similarly, for areas that need vehicular transportation, it may be expected that awareness spaces will follow any major transportation arteries. Another feature that may have an effect are edges (Brantingham and Brantingham, 1993b) which may be physical or psychological boundaries. One example would be the borders of non-porous neighbourhoods, where strangers are often present and so are accepted. However, travelling beyond them and into the neighbourhoods may be curtailed (and so would their awareness spaces and potential offences) because they may arouse suspicion and be challenged by residents.

The environmental backcloth will also influence where an offender’s awareness space results in an offence by way of its convergence with suitable targets. For example, consider the original scenario of an offender who prefers to offend in their awareness space, it assumes this can be realised anywhere as potential targets are homogenous and uniformly
distributed. This, however, is not true. Targets are heterogeneous and unevenly distributed in space and time. As illustrated for the spatial dimension in Figure 2.4, their distribution will interact with an offender’s awareness space to the extent that their offences will vary with the concentration of (suitable) targets. Certain targets or concentrations of targets will also bias an offender’s awareness space and their target searches. As noted by Brantingham and Brantingham (1995), these include:

- *crime generators*, such as shopping areas, which produce crime because they attract large numbers of people (including offenders) for reasons unrelated to crime and so create opportunities;
- *crime attractors*, such as drug markets, which produce crime because they specifically attract criminals due to the opportunities they present; and
- *crime neutral areas* that produce little crime because they attract neither criminals nor non-criminals. Clarke and Eck (2003) also added crime enablers, such as some recreation parks, which produce crime because there is little regulation of behaviour.

**Figure 2.4:** The likely distribution of crime due to the convergence of awareness space and areas with many potential targets. Adapted from Brantingham and Brantingham (1981b).
Beyond this, the crime pattern perspective of crime also emphasises the non-stationarity of the environmental backcloth. This can be in the short-term where, for example, locations (including the types outlined above) are only suitable for an offence or have the same type of effect at certain times of the day or days of the week. An example would be a bar that is closed in the morning and as it does not attract any custom, it acts as a crime neutral location at that time. When it opens later in the day, it may attract many patrons and so produces lots of opportunities and acts as a crime generator. In the evening it may be infamous for the types of opportunities present and so acts as a crime attractor, and after closing and all the managers leave it may act as a crime enabler as there are no deterrents to offending. The backcloth may vary in the mid-term such as when a school or university re-opens for the new academic year and changes people’s routines, and admits a new group, of potential offenders and victims. Finally, its effects may also vary in the long term such as when an offender ages and through experience or greater access to vehicles or public transport develops a larger awareness space; or when they change addresses and develop new (and forget old) awareness spaces. Long-term migration in work, shopping and leisure locations to fringe areas of the city will also cause widespread changes in activity spaces (and awareness spaces) of potential victims and offenders.

2.2.4. Social disorganisation and collective efficacy perspectives

Although arguably distinct, the social disorganisation (Shaw and McKay, 1942) and collective efficacy (Sampson, Raudenbush and Earls, 1997) perspectives can be considered refinements of the same idea. They share a common genealogy in the work from the Chicago School in the 1920s and particularly the research by Park, Burgess and McKenzie (1925/1967) on the urban ecosystem and the natural arrangement of the city around the central business district. That is, as illustrated in Figure 2.5, the city may be divided into concentric zones where each possesses relatively homogenous socio-
structures but increase in general prosperity the further from the central business district. Although its possible relationship with crime was briefly mentioned, for example in the chapter by Burgess (1925/1967) and relatedly in other work such as that by Sutherland et al. (1939/1992), it was not until Shaw and McKay (1942) it became a focus.

In their article, Shaw and McKay (1942) propose that variations in the delinquency rate (as measured using the home locations of delinquents) will be related to the ability of communities to establish common values and maintain informal social control (through social organisation) such as by intervening against suspicious activities. Although they also considered other variables, they particularly hypothesized and found results consistent with delinquency being related to poverty, transiency, and ethnic heterogeneity. They tentatively suggested several mechanisms for these. In terms of poverty, they suggested

**Figure 2.5: The concentric zone model of urban socio-structures. Adapted from Park et al. (1925/1967).**
an indirect influence where those in disadvantaged areas are more exposed to social norms that may condone delinquency. In addition, in areas with greater affluence, residents are more likely to own their homes and be permanent residents and so are unlike transient residents who may not inhabit a location long enough or be willing to expend effort to decrease delinquency. Regarding ethnic heterogeneity, they suggest that residents may be unable to communicate, or they hold discordant social values preventing the formation of social organisation. They also note that these spatial patterns show long-term stability despite changes in residents suggesting the observations are endemic to the areas and not to the populations.

Although seminal, later interest in social disorganisation and related perspectives was intermittent until the 1980s and 1990s. Here, papers such as those by Sampson (1985) and Sampson and Groves (1989) proposed and tested a number of additional indicators of social disorganisation. These included housing and population density where the greater number of people inhabiting an area, the more difficult it is to recognise neighbours; and family organisation where areas with substantial amounts of lone-parent households are unlikely to provide the same amount of supervision as dual-parent households. At this time, Sampson and colleagues (e.g. 1997) also introduced the concept of collective efficacy which regards the shared ability and willingness of communities to activate social control. The key difference from classical social disorganisation is that members of the same community not only have relationships with other members, but they have mutual trust and support to act in the best interest of the community. That is, while previous examinations of the relationship between social ties and social control was largely assumed (see Sampson and Groves, 1989); Sampson et al. (1997) argued and found collective efficacy to be a mediating mechanism between neighbourhood structure and crime (which in that study concerned various measures of violence).
2.2.5. Eyes on the street perspective

The *eye on the street* perspective, also known as new urbanism and the encounter model is commonly attributed to the writings of Jane Jacobs and particularly her 1961 book, *The Death and Life of Great American Cities*. Based on her anecdotal observations of neighbourhoods, mostly in New York, Jacobs suggested that for most streets, public order is not predominantly maintained by the police, but informally and unconsciously by the network of people who use it. This is to the extent that to reduce crime, a proscription from the theory is that a street should be permeable, for example by having a grid layout, and be well used by residents and/or strangers (non-residents). The street will therefore have many guardians and an amount of natural surveillance that will inhibit crime and ensure safety. She also argued that (residential) streets should be of mixed-use and contain stores and other types of public places – particularly those primarily used during the day-time and night-time. This is for several reasons. First, and because of the need for general surveillance, these places will attract and give non-residents (as extra eyes) reasons for being on the street. By being in use during the daytime (when residents will likely be away at work) and night-time (when residents are likely in their homes), these public places will help provide a continuous flow of surveillance and *eyes on the street*. Third, because the storekeepers (and alike) are themselves important providers of natural surveillance, and by being invested in their business, they will additionally likely be active in ensuring peace and order.

2.2.6. Defensible space perspective

Although often viewed as the antithesis of the eyes on the street perspective, Oscar Newman drew upon this perspective while developing his defensible space perspective (also called the enclosure model) in the book of the same name in 1972 (see also Coleman, 1985; Newman, 1996). Based on this and his crude analysis of crime rates in several housing projects in New York, he identifies three major factors associated with crime. The first is
anonymity. The idea was that in high-density buildings, areas and those that are shared or used by countless people, or areas that offer easy access to the public and are used as a through-path, it becomes impossible to recognise residents from intruders. They will therefore be unable or less able to recognise suspicious circumstances and intervene. Relatedly, the second factor regards personal responsibility. Here, the idea is that in areas or buildings that are shared or used by many people, the residents may feel little association or responsibility for the grounds, including shared areas such as hallways. As such, they will likely make little effort to maintain or protect those areas or participate in surveillance.

The final factor is that of permeability where in housing projects that have a multitude of (physically or psychologically) uncontrolled entry and exit points, it not only becomes easier (and more justified) to use it as a through path (see before), it also makes it simpler for offenders to enter and exit.

It is from these observations that Newman (1972) coined the term territoriality, in terms of the creation of perceived zones of (territorial) influence, as the cornerstone of his defensible space perspective. As illustrated in Figure 2.6, Newman contended that the demarcation of space into a hierarchy of public and private (and graduations thereof) zones using physical and psychological barriers disrupts the movement between. It follows that on the one hand, the controllers (residents) of the private zones are then able (and encouraged) to take ownership of their zone and consider them as part of their personal space. From this, Newman argued that the residents will then take some personal responsibility and by setting up a code of behaviour (potentially with other residents); they will be willing and able to detect and intervene in suspicious circumstances to protect their space. For example, if an intruder entered. On the other hand, the demarcation of public and private spaces and the subsequent changes in residents’ behaviour should discourage intruders and offenders from trespassing. In turn, further reducing anonymity and making
it easier for residents to be vigilant and control *their* space. In effect, this creation of territory should reduce anonymity, increase personal responsibility, and create (theoretically at-least) impermeable environments.

### 2.3. Theoretical model of burglary

Taking these theories together, a hypothetical model of the factors expected to influence burglary location choices can be theorised (that will be later tested in Chapters 10 and 11). Although these factors can be thematically divided in a number of ways, one way is if they regard the offender and their movement and the road network; passers-by and their movement and the road network, and other factors.

#### 2.3.1. Offenders and the road network

Based on rational choice and the notion that offenders, including burglars, attempt to minimize effort when deciding where to offend, it would be expected that offenders will
attempt to minimize the amount of travel to offend. In fact, in ethnographic studies offenders consistently report that distance is an important factor in offending decisions (e.g. Brown and Altman, 1982; Rengert and Wasilchick, 1985). Similar results are obtained from quantitative studies that have investigated the distance between the offender’s home and offence locations (e.g. Snook, 2004; Wiles and Costello, 2000). However, in most studies, distance is usually measured in terms of Euclidean or the straight-line distance to the offence location (e.g. Clare, Fernandez and Morgan, 2009). While this will provide a good estimate of the likely travel required, it is clearly imperfect. The amount of travel required will obviously be linked to the shape of the road network and how easy or direct it is to travel to the location. As such, the distance along the road network would be expected to be negatively associated with a location being selected for a burglary.

The common finding that the frequency of offending is inversely proportional to the distance offenders must travel is also expected to emerge from the crime pattern perspective. As described earlier, through their daily activities offenders develop an awareness of criminal opportunities. Their daily activities will however be most likely centred around their key activity node(s) such as their home. Based again on the idea that offenders, like non-offenders, will try to minimise effort, their offending and non-offending activity will also generally decrease further from their home. As a result, they are more likely to be familiar with locations closer to their home. However, while previous researchers have generally assumed that distance provides a reasonable estimate of a location’s familiarity (and effort), it is likely more nuanced than this. Offenders will have additional activity nodes, such as nearby shopping areas. The street configuration will also play an important role as it will determine how likely or often offenders will travel along a particular street to reach those or other locations. For example, some streets such as major roads will, on average, feature on many journeys and so will be more familiar than
others such as cul-de-sacs which will feature in fewer journeys. As such, and independent of the effect of distance, because offending in known locations reduces the uncertainty about the likely outcome and also reduces any effort needed to scout that area to reduce this uncertainty, it would be expected they would be more likely to offend in more familiar locations.

2.3.2. Passer-by movement

Based on the theories described above, passers-by or people going about their daily activities who have the potential to provide ambient guardianship (Felson, 1998) or regulate behaviour (Miethe and Meier, 1994) will also be expected to influence burglary decisions. That said, there is no general consensus in the theories or related research regarding this effect. To explain, recall that Jacobs (1961) suggested that guardianship is provided by everyone (not engaged in criminal activity) present on a street, as their presence provides ‘eyes on the street’ and ‘natural surveillance’ that can deter crime. In contrast, Newman (1972; see also Coleman, 1985) argues that whether the ambient population deters crime depends on who is present. That is, informal guardianship is expected to only be provided by those who are residents (or local people) and only in places where strangers (non-local people) can be identified. Similarly, the theories by Felson (1995) and later Reynald (2011) suggest that a person’s likely capability as a guardian and how much they will deter crime will depend on their attachment or proximity to an area.

While these theories are not unreasonable, the findings from quantitative research intended to test them is mixed and can be broadly divided into two branches of research. The first, which includes older area-level (White, 1990) and more recent street-level studies (e.g. Armitage, 2007; Beavon, Brantingham and Brantingham, 1994; Johnson and Bowers, 2010; Hakim, Rengert and Shachmurove, 2000) generally suggests that areas
which are expected to have greater overall levels of movement are more likely to experience crime. In contrast, the second branch of research, which utilise the space syntax methodology to be discussed in Chapter 3, generally find, as per Jacobs’ (1961) ideas that locations which are expected to have greater amounts of movement have less crime (e.g. Hillier, 2004; Chih-Feng Shu, 2000; see also Nubani and Wineman, 2005 for an exception). Others also find that streets which are characterised as being more likely to be used for localized movement tend to experience less crime (Hillier and Sahbaz, 2008). It should be noted that the second branch of research tends to be weaker in terms of the statistical methods employed. For example, they tend to rely on descriptive statistics (Chih-Feng Shu, 2000) or fail to account for other variables known to influence crime (Hillier, 2004; Hillier and Sahbaz, 2008). As such, and while its precise effect is currently debateable and open to further research, passer-by movement, either in terms of overall volume or the volume of the different types of people (local or non-local people) is expected to influence offending decisions.

2.3.3. Other factors

Along with these influences, the theories described in Section 2.2 also highlight other factors that are expected to influence offender decision-making. First, and according to the rational choice perspective, offenders weigh up the potential rewards from an offence, burglars would be expected to prefer more affluent dwellings where greater proceeds are expected. While this is supported by research which suggests burglars are financially motivated (e.g. Maguire and Bennett, 1982; Rengert and Wasilchick, 1985), conversely, more affluent dwellings would be expected to have better security measures (e.g. burglary alarms). Therefore, and calling upon the rational choice perspective again, they may be riskier or may be more difficult to break into and so require more effort. More affluent properties would be therefore less likely to be targeted.
Burglary location choices are also likely influenced by the levels of social disorganisation (Shaw and McKay, 1942), social cohesion and related concepts in a neighbourhood (e.g. Sampson and Groves, 1989; Morenoff, Sampson and Raudenbush, 2001). That is in terms of how closely knit the neighbourhood is and so how willing residents may be to act collectively as guardians against crime which has been shown to play a role in neighbourhood crime rates (e.g. Bruinsma et al., 2013).
Configurational Methods
In Chapter 2, the importance of movement - including that of passers-by in terms of being potential guardians - in the crime event was shown. In the current section, the current approaches for counting or estimating movement are introduced and discussed. It is then argued that configurational network methods, as developed and used in the fields of graph theory and space syntax, are the most promising method for estimating movement. These are examined in further detail and the relevant terminology and technical material for these methods are also introduced.

3.1. Introduction

As already identified, the ability to measure movement flows has a key role in testing many of the prevailing perspectives in environmental criminology. That being said, given the costs, the other resources needed, and the additional problems with automated (e.g. CLIP, 2007) and manual (e.g. May, Hopkinson and Turvey, 1991) counts of movement, the current focus in related research are methods for estimating movement. Although some methods, such as simulations or agent-based models are not well established for
modelling larger urban areas (e.g. Clifton et al., 2004) and so are not discussed, these methods can be divided into four types: perception, land-use, road type and configurational methods.

### 3.2. Perception methods

Perception-based methods are those that attempt to estimate movement flows, typically on an ordinal scale (e.g. none, light, moderate or heavy traffic), through surveying the perceptions of traffic. That is, without any direct attempt to count pedestrians or vehicles. This method has been used in previous criminological research in two ways. The first is through surveying independent observers (e.g. researchers), for example, where Loukaitou-Sideris et al. (2001) had two researchers visit streets to estimate vehicular and pedestrian traffic on a three point scale (light, moderate or heavy). Likewise in Luedtke (1970) where surveys of pedestrian traffic were conducted by field staff and in Harold Lewis Malt Associates (1973) by police officers.

The surveying can also be through non-independent observers (e.g. residents). For example, in Hunter and Baumer (1982) who interviewed 556 adult residents and asked them on a four point scale how many people during the daytime and after-dark would be “…usually on the street in front of [their] house?”. The same question was also asked regarding vehicular and pedestrian traffic in Fowler Jr (1979) and regarding students in Dietrick (1977). In addition to these, there are other studies that try to estimate traffic levels through monitoring traces of presence such as parked vehicles being present and/or children’s toys in yards (Brown and Altman, 1983; Pablant and Baxter, 1975).

Although these methods are now generally unused, it is still important to consider their advantages and disadvantages. The major advantage, as suggested in Hunter and Baumer (1982) and Loukaitou-Sideris et al. (2001), is that the estimates can be easier and quicker
to collect than actual counts. This will be particularly true when using non-independent observers such as residents who can be contacted remotely, for example, by phone in Fowler Jr (1979). A key disadvantage as discussed in Motoyama et al. (1980) is that whilst behaviour, such as offending, will likely follow (the offender’s) subjective perceptions of traffic rather than actual counts, the offender’s and the observer’s perception may not match. This is supported in terms of independent observers where Loukaitou-Sideris et al. (2001) dropped their measures from their analyses as the two researcher’s data “was proved to be too subjective” (p.11). The offender and the observer may also differ in their ability to assess traffic where Fowler Jr (1979) found that actual changes in traffic counts were relatively poorly reflected in their perceptions data (see Table 3.1) whereas Newman (1996) found in their study that a (36%) decrease in traffic was recognised by 73% of residents with only 14% thinking there was an increase in traffic (see also City of Dayton, 1994).

3.3. Land use methods

The second type are land-use methods, also called sketch plan methods (Clifton et al., 2004), attempt to estimate or approximate movement flows as a product of the presence (or absence) of various land uses. Although not explicitly used within criminology, parallels can be drawn to studies of crime generators (Brantingham and Brantingham, 1995; see

Table 3.1: The observed and perceived changes in vehicular traffic between 1976 and 1977. Adapted from Fowler Jr (1979).

<table>
<thead>
<tr>
<th>Change in Traffic (Actual Count)</th>
<th>Perceived Change in Traffic (Survey)(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Heavier(^1)</td>
</tr>
<tr>
<td>Blocked Streets</td>
<td>-75%</td>
</tr>
<tr>
<td>Narrowed Streets</td>
<td>-17%</td>
</tr>
<tr>
<td>Untreated Streets</td>
<td>+5%</td>
</tr>
</tbody>
</table>

\(^1\) Percentage of respondents

\(^2\) Based on the question in the 1977-only survey: “…how about the cars, motorcycles, and buses that pass in front of your home during the day – would you say the traffic is heavier than it was a year ago, lighter, or about the same?”
Environmental criminology) in terms of the land uses which tend to be analysed are trip generators. These analyses in effect assume traffic volume corresponds to the presence of various types of premises (e.g. Anderson *et al.*, 2012; Davison and Smith, 2003; LaGrange, 1999; McCord *et al.*, 2007; Pablant and Baxter, 1975). That is where presence can be measured in various ways from being present in the unit (Davison and Smith, 2003) or nearby (Pablant and Baxter, 1975) to more sophisticated measures such as their proximity (Anderson *et al.*, 2012) or their (kernel-estimated) density (McCord *et al.*, 2007).

Outside of criminology, these methods have been far more explicitly used to model pedestrian and vehicular volumes including being evaluated against actual count data (see Table 3.2 and also Appendix A for a comprehensive list). This is done by collecting data on a range of variables at and around count sites which are expected to influence movement flows. Multivariate regressions are then conducted to calculate their relationship with the observed count data which can then be used to estimate counts where actual count data is not collected. The first of these studies, by Pushkarev and Zupan (1975) and Behnam and Patel (1977), attempted to predict pedestrian volumes (almost exclusively) using the surface areas of different types of land-uses at the count site and achieved modest adjusted $r^2$ values (0.20-0.60). Later studies incorporated a greater variety of variables including those relating to the infrastructure itself which may attract or deter traffic such as the presence of a sidewalk or the width of the adjacent road (e.g. Qin and Ivan, 2001). They also included variables related to demographics such as residential and employment populations or densities (e.g. Schneider, Arnold and Ragland, 2009) and the presence of facilities such as metro stations and schools (e.g. Schneider, Arnold and Ragland, 2009).

Whilst there are limited diagnostics of the models (see also below) in the land-use studies reviewed (see Table 3.2), certain variables and models can be identified which appear to
correspond to models which better fit the data (summarised in Table 3.2). In terms of the pedestrian volume prediction models, they tended to include variables about the location of high activity areas, transit stops, universities/schools and measures of the employment/commercial and population/household density. These models tended to achieve $r^2$ values of around 0.80. In comparison, the vehicular prediction models tended to incorporate variables concerning the type of road (e.g. number of lanes) and its accessibility to a highway along with measures concerning the presence of a high activity area and population/household and employment/commercial densities. These also tended to achieve $r^2$ values of around 0.80.

Many of these studies however have a number of weaknesses. For one, and as said above, they tend to have limited diagnostics of their models. Many tend to over-rely on $r^2$ values

---

**Table 3.2: The four best fitting land-use models (and the variables used within each model) for predicting pedestrian and vehicular traffic.**

<table>
<thead>
<tr>
<th>Pedestrian Models</th>
<th>Vehicular Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access to highway</td>
<td>● ● ● ●</td>
</tr>
<tr>
<td>Bus routes/frequency</td>
<td>●</td>
</tr>
<tr>
<td>CBD or other high activity area</td>
<td>● ●</td>
</tr>
<tr>
<td>Employment/commercial density</td>
<td>● ● ●</td>
</tr>
<tr>
<td>Land use mix</td>
<td>● ● ●</td>
</tr>
<tr>
<td>Population/household density</td>
<td>● ● ●</td>
</tr>
<tr>
<td>Road type (Inc. number of lanes)</td>
<td>●</td>
</tr>
<tr>
<td>Slope/gradient</td>
<td>●</td>
</tr>
<tr>
<td>Street/path length</td>
<td>●</td>
</tr>
<tr>
<td>Transit stops / stations</td>
<td>●</td>
</tr>
<tr>
<td>University campus / schools</td>
<td>●</td>
</tr>
<tr>
<td>Other</td>
<td>●</td>
</tr>
<tr>
<td>Adjusted $r^2$ (or equivalent) (x100)</td>
<td>80 90 80 79</td>
</tr>
</tbody>
</table>

The land-use models are from the following papers: 1 Qin and Ivan (2001); 2 Schneider, Arnold and Ragland (2009); 3 Schneider *et al.* (2013); 4 Tabeshian and Kattan (2014); 5 Mohamad *et al.* (1998); 6 Zhao and Chung (2001); 7 Zhao and park (2004); 8 Anderson, Sharfi and Gholston (2006).
for assessing fits and comparing alternative models - even when the numbers of regressors vary. Few papers also conduct cross-validation exercises that could reveal over-fitting. Very few also evaluate against models from other papers (for the exception see Yang, Wang and Bao, 2014) which would not only allow comparisons between established alternative models but would also allow the generalizability of the models (and their estimated coefficients) to be assessed. As it stands, little is known about the efficacy of most of the models (particularly) outside of the original study area and sample.

### 3.4. Road type methods

The third type of methods for predicting traffic are road type-based and these estimate movement based on the type of road or types of roads nearby. To start with the more basic of these, White (1990) when examining burglary risk suggested that certain types of roads, categorically-defined arterials (or main roads), will contain greater volumes of (vehicular) traffic. From this, and at the areal level, they considered that neighbourhoods having more roads directly connected to these arterials will have “greater opportunity for movement” (p.61) and therefore greater (vehicular) traffic. In a similar way, Greenberg, Rohe, and Williams (1982) and Greenberg and Rohe (1984) both considered that US census blocks bounded by major thoroughfares (four- or six-lane arterial roads), for which they reference Gardiner (1978) in describing them as “movement generators” (p.37), will also experience greater traffic. This approach is also used in Pablant and Baxter (1975) who in their assessments of a school’s observed activity level considered if the school grounds were delimited by commercial streets.

Instead of categorical definitions of major and minor roads, other papers have used hierarchical ordinal street classifications. In Beavon, Brantingham, and Brantingham’s study in Canada this was according to a “classification scheme used by city planners” (1994, p.126) where roads were defined as feeders, minor arteries, major arteries or highways.
This was similarly done in Davison and Smith’s US study albeit they included little detail on each category except that the lowest included “less travelled residential roads” and the highest “represent[ed] major arteries” (Davison and Smith, 2003, ‘Accessibility Measures’ section, para. 2). In the (only) equivalent UK study, Johnson and Bowers (2010) used the Ordnance Survey’s official classification to designate street segments as major roads, minor roads, local roads, or private roads (as well as manually coding cul-de-sacs to contrast with through roads) based on their intended type and level of use. This paper also includes the type of road each segment is directly connected to which may also affect its volume of traffic. In addition, although this is the only paper that specifically considers this first-order connectivity, others have incorporated similar measures such as proximity to certain types of roads or other features. In particular, Hakim, Rengert, and Shachmurove (2000, 2001) considered proximity to highway exits whereas Dietrick (1977) used separate proxies for vehicular traffic: the number of blocks to the nearest freeway exit and the city’s entertainment complex; and the number of blocks to the nearest school and park for pedestrian traffic.

Although the network methods presented so far are relatively simplistic, this is a key advantage as they can be calculated for large study areas quickly and are easy to interpret. They also, perhaps surprisingly, correspond reasonably well with observed traffic counts (see Table 3.3). Penn et al. (1998) found route hierarchy in London (UK) correlated reasonably well \( r = 0.81 \) with vehicular flow. Additionally, Zhao and Chung’s study in Broward County (FL) found road classification \( r = 0.84 \) and number of lanes \( r = 0.77 \) correlated well with vehicular traffic whereas being directly connected to an expressway \( r = 0.48 \) and the distance to an expressway exit \( r = 0.20 \) did less so (2001). This is also shown with the mean vehicle counts on the different road types from Zhao and Chung (2001) and Lowry’s study in Moscow (ID) (2014) (see Table 3.3). On the other hand, and
as also shown in Table 3.3, in the study in the City of Minneapolis (MN) by Lindsey et al. (2012), road classification does not appear to correlate as well for pedestrian and bicycle traffic, though there still somewhat tends to be greater traffic on the more prominent road types. This is clearer when the main road type (principal arterials) are ignored, as they may not be suitable for these modes of travel.

### Table 3.3: The observed mean hourly numbers of vehicles and pedestrians for different classes of roads.

<table>
<thead>
<tr>
<th></th>
<th>Vehicular</th>
<th></th>
<th>Pedestrian</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Counts$^1$</td>
<td>Mean$^1$</td>
<td>Counts$^2$</td>
<td>Mean$^2$</td>
</tr>
<tr>
<td>Principal Arterial</td>
<td>184</td>
<td>142</td>
<td>37</td>
<td>504</td>
</tr>
<tr>
<td>A-Minor</td>
<td>278$^4$</td>
<td>92$^4$</td>
<td>47$^4$</td>
<td>152$^4$</td>
</tr>
<tr>
<td>B-Minor</td>
<td></td>
<td></td>
<td>75</td>
<td>39</td>
</tr>
<tr>
<td>Collector</td>
<td>305</td>
<td>42</td>
<td>107</td>
<td>157</td>
</tr>
<tr>
<td>Local</td>
<td>49</td>
<td>25</td>
<td>150</td>
<td>32</td>
</tr>
</tbody>
</table>

1 Source: Zhao and Chung (2001); counts based on automated pneumatic-tube (axle) counters.
2 Source: Lowry (2014); counts based on automated pneumatic-tube (axle) counters.
3 Source: Lindsey et al. (2012); counts based on field observations (for pedestrian and bicycle traffic) and also magnetic loop detector counts (for bicycle traffic).
4 Separate figures are not presented for these as they were aggregated into minor arterial roads in their respective papers.

In addition to these relatively simplistic road type-based methods, there is also a related family of sophisticated configurational methods. These methods were developed in the fields of graph theory and space syntax and function by abstractly representing the street and road networks as a graph of interconnected spaces. Those graphs are then analysed using configurational measures to estimate the parts of the graph which are most likely to experience movement.

Owing somewhat to the work from space syntax, there have been at least 78 studies (see Appendix B) assessing these methods and are often cited as explaining between 60% and
80% of the variance in movement flows (e.g. Hillier and Vaughan, 2007; Penn, 2003). While this performance is similar to that achieved by the land-use methods (but likely better than the other competing methods), configurational methods have several other advantages. Foremost, unlike the land-use methods, configurational methods have been replicated, applied, and tested in a large number of cities and countries around the world (see above). They are also already relatively frequently, and successfully, used in analyses of crime (Johnson and Bowers, 2010; Davies and Johnson, 2015; Frith, Johnson and Fry, 2017). A final advantage is that how these methods are calculated and can be theoretically justified, their modification and application to the understanding of individual movement (see Chapter 5) is relatively straightforward.

These methods are therefore used in this thesis. However, these configurational methods can be computed in many ways including how they convert and represent the street network as a graph, what measure they use to analyse the resulting graph, and what parameters they use with the measures. These are now discussed.

3.5.1. Representations

Prior to analysing the street network, it must be transformed into a graph and represented mathematically as a collection of vertices \( V \) and edges \( E \) which represent the connections between vertices. This translation of the street network into a graph can occur in two main ways. The first and most direct way, but relatively uncommon in the literature, is the \textit{primal representation}. Here, zero-dimensional cartographic entities (e.g. street junctions) are represented by their equivalent graph entities (vertices) and one-dimensional cartographic entities (e.g. segments between junctions) by theirs (edges) (see also Porta, Crucitti and Latora, 2006b). In the second and often described as the indirect way, the \textit{dual representation}, this is reversed where the connections between junctions (e.g. street segments) are represented by vertices and the junctions (where street segments intersect...
and connect with each other) by edges (see also Porta, Crucitti and Latora, 2006a). These, and the difference and translation between the two, are shown below in Figure 3.1. Whilst the former may be more intuitive and retains the street network’s morphological appearance, the latter is favoured for a number of reasons. For one, the dual representation is said to represent the information space of the network in terms of how it is articulated and navigated (Masucci, Stanilov and Batty, 2014; Rosvall et al., 2005). For example, how people navigate from segment to segment rather than junction to junction. It is also preferred as the configurational measures typically analyse vertices (Crucitti,

Figure 3.1: Maps illustrating the construction of primary and dual graphs (using street segments) of the street network of a village (near Chesterfield, UK).

Maps showing the original road network (a), the primary graph where vertices have been added at junctions and edges have been added between vertices (b), where edges are represented as vertices (c) and the dual graph where the junctions have been replaced by edges (d).
Latora and Porta, 2006) where for example, when modelling traffic, it is the traffic along a street or street segment (represented as vertices in dual graphs) that is usually of interest rather than that at a junction.

Given dual representation, the literature uses a number of different units from which to construct the vertices (see Figure 3.2). The most fundamental of these and what could be considered the standard graph theoretic definition is that of street segments. These are formed using street centreline data (see also Dhanani et al., 2012; Turner, 2007) by taking sections of the street network between two junctions or between a junction and the street’s end-point (see Figure 3.2a). That is, between locations where there is a choice of direction. These are advantageous as they are at the resolution at which the network is traversed in that navigation is from segment to segment and each is traversed in its entirety if it is on the path (except at origins and destinations). That said, they are not likely to be perceptually discrete which limits the plausibility of using certain impedances (see below).

Given this and how some research suggests how routes may be chunked and adjudged by the number of cognitively-distinct sections (Allen, 1981; Allen and Kirasic, 1985), there are other representations, including those that aggregate street segments, based on some continuity to build what are likely to be cognitively distinct units. The simplest is semantic continuity where street segments are aggregated, and represented as vertices, if they share the same street name to form named streets (Jiang and Claramunt, 2004; see Figure 3.2b). Often this aggregation ignores street name qualifiers, such as ‘upper’ and ‘lower’ or ‘north’ and ‘south’, although the reverse is possible. Also, and while this representation likely incorporates some cognitive aspect of the urban environment that we use to navigate, it is not without problems. For example, as noted by Jiang et al. (Jiang, Zhao and Yin, 2008), there are difficulties when data, such as street names, are incomplete. Also, when including path data, the paths may not be named or otherwise uniquely referenced.
Figure 3.2: The street network of a village (from Figure 3.1) represented with different units.

A) Street segments, B) Named streets, C) Axial lines, D) Minimal fewest lines (DepthMapX using Ordnance Survey road casings), E) Axial segments (with stubs removed), F) Self-organised natural streets (45°)
Vertices can also be defined from the street network using visible or angular continuity. Although not formed from street segments themselves, the most popular case of this are from space syntax and the work beginning from Hillier and Hanson (1984). The first of these, axial lines, are hand-drawn to break up continuous open spaces into the fewest and longest straight lines that pass through every accessible space (Hillier and Hanson, 1984; see Figure 3.2c). They can be thought of as lines of sight (Penn et al., 1998) where each line is a vista covering the area that can perceived from a single point of view (Jiang, Claramunt and Klarqvist, 2000). Being hand-drawn, however, is a major point of contention as the representation must be arbitrary where the same system can plausibly be drawn differently by different people and even by the same person at different times (e.g. Batty and Rana, 2002). These slight differences, and they only need to be slight (Ratti, 2004b; a; see also Hillier and Penn, 2004), can result in different systems and results.

Whilst there is this subjectivity in hand-drawn axial maps, attempts have been made to automate their construction. For example, by the introduction of the ‘all-line map’ which builds lines of sight between inter-visible vertices (Penn et al., 1997). Using greedy algorithms, the map can then be reduced to ‘near-minimal sets’ of ‘fewest lines’ and a further reduced ‘minimal fewest lines where extra depth minimising lines are removed (Turner, 2004; see Figure 3.2d). Although these are again, and problematically, not necessarily unique (Batty and Rana, 2002). Another key problem with all types of axial lines, particularly as they can be of unlimited length (Dalton, 2001), is they can aggregately obscure any differences in syntactic values and people counts along the line. To overcome this latter problem, and now the predominant approach within space syntax, axial segments were introduced. For this representation, they take the original axial lines but break them at intersections into the axial segments which are then treated as vertices in graphs (Dalton, 2001; Turner, 2001; see Figure 3.2e).
This Gestalt principle of visible continuity can also be applied to aggregating street segments into cognitively meaningful units. The resulting units being generally referred to as ‘self-organised natural streets’ (Jiang and Liu, 2009; Thomson, 2004; Figueiredo and Amorim, 2005). In this process, the street segments are combined with neighbouring ones if their deflection angle is below a pre-set threshold (e.g. 45°) (see Figure 3.2f). This combining can occur using three different strategies, but the one that yields unique sets, is the ‘self-best-fit’ strategy (Jiang, Zhao and Yin, 2008) In this, every combination of street segments meeting at a junction are ranked based on their deflection angle. The combinations with the smallest angles (below the threshold) that are not involved in other combinations with smaller angles (at that junction) are then joined.

3.5.2. Measures

Given these representations, the resulting graphs are generally analysed using variations of four key measures. Six of these using the street segment representation (see earlier) and angular impedance (see later) are shown in Figure 3.3. Starting with the simplest and most local because it only looks at immediately neighbouring vertices, is ‘degree’ (or ‘connectivity’ as it is known within space syntax) (see Figure 3.3a). This is calculated as the number of edges incident to a vertex and can be denoted for vertex $e$ by:

$$\text{degree}_e = N_{\text{edges}_e}$$  \hspace{1cm} (1)

A related measure, and one which is only used within space syntax, is that of ‘control’ (see Figure 3.3b). This is described as a measure of how much a vertex controls or offers access to and from its neighbours (Hillier et al., 1993). It is calculated as the sum of the reciprocals of ‘degree’ values of its directly connected vertices (x) such that for vertex $e$:

$$\text{control}_e = \sum_{x \in N_{\text{edges}_e}} \frac{1}{\text{degree}_x}$$  \hspace{1cm} (2)
Figure 3.3: The street network of a town (Chesterfield, UK) analysed with six different measures.

A) Connectivity; B) Control; C) Closeness (no radius); D) Betweenness (no radius); E) Closeness (1000m radius); F) Betweenness (1000m radius). The measure scores are coloured from green (low) to red (high).
There are two other measures that are popular within the literature. The first, ‘closeness’, which when normalised in space syntax is called ‘integration’ (see later), is calculated for each vertex as the inverse of the mean shortest distances from itself (the focal vertex) to all other vertices (Freeman, 1979) (see Figure 3.3c). That is if the shortest distance between the focal vertex \( e \) and all other vertices \( j \) is denoted by \( d_{ej} \) then closeness can be calculated by:

\[
closeness_e = \left[ \frac{\sum_{e \neq j \in V} d_{ej}}{n} \right]^{-1}
\]

(3)

Closeness can also be interpreted as the accessibility of a vertex, or in space syntax terms the ‘to-movement’ potential of a vertex (Hillier, 2005), in that a space that is closer to more spaces is more likely to be visited (and therefore experience more movement or traffic). On the other hand, ‘betweenness’, also called ‘choice’ when normalised in space syntax (see later), is calculated for each vertex as the number of (shortest) paths from all vertices to all others which pass through it (Freeman, 1977) (see Figure 3.3d). Formally this is defined as:

\[
betweenness_e = \sum_{i,j \in V, i \sim j} \frac{\sigma_{ij}(e)}{\sigma_{ij}}
\]

(4)

Where \( \sigma_{ij} \) denotes the number of shortest paths between vertices \( i \) and \( j \) and \( \sigma_{ij}(e) \) denotes the number of those that pass through \( e \). Betweenness is interpreted in space syntax as the ‘through-movement’ potential of a vertex (Hillier, 2005) and can be likened to a simplistic model of how people move around the city, approximately from everywhere to everywhere else, using the shortest routes; and where these routes overlap, a greater volume of co-presence is expected. It should also be noted that the different
implementations of each metric in different software packages can result in slightly different metric scores and analyses (see Varoudis et al., 2013).

Closeness and betweenness can also be calculated at a localised scale and with a weighting scheme. In terms of the former, this is where the current measures quantify each vertex in terms of the entire network. However, some types of movement, for example pedestrian, likely are not influenced by distant configurations. To recognise this, a radius may be used that excludes vertices from the metric calculation that are a certain distance away from the focal vertex \(e\). That is if the radius is denoted by \(r\) then a condition of \(d_{ej} \leq r\) (for closeness) and \(d_{ij} \leq r\) (for betweenness) is supplied. For closeness, this limits it to calculating the inverse of the average distance to vertices that are at most \(r\) away. Similarly, with betweenness where the routes calculated from each vertex is limited to those that are at most \(r\) away. Although space syntax tends to set these radii arbitrarily to optimise empirical results (e.g. Hillier and Iida, 2005), a theoretical justification could be to model realistic journey lengths where, for example, only destinations up to 1km are used to model feasible non-vehicular (pedestrian) journeys (see Figure 3.3e for closeness and Figure 3.3f for betweenness with this radius; see also impedance).

Weighting can also be applied to account for non-equivalent vertices. For example, and considering betweenness, it would be expected that there are more journeys to and from vertices with large numbers of buildings vertices with few or none. Although rare in published analyses, various types of weighting can be used including its street length (e.g. Chiaradia, 2007; Turner, 2007 see also Chapter 5).

**3.5.3. Impedances**

Calculating these measures and setting radii requires stipulating some measure of impedance or distance. That is, many of the measures (e.g. closeness and betweenness)
rely on defining ‘shortest paths’, but as a longstanding topic in the literature, there is
debate on how distance is conceptualised. Also, and although there is some support for
certain wayfinding logics, their use can be problematic in conjunction with some
representations. Nonetheless, and although Golledge and Garling (2002) list 24 types of
possible route selection criteria, the literature predominantly use three types of
impedances. These are illustrated in Figure 3.4 where even in the small network, the
different impedances (and representations) can result in different shortest paths.

Originally, space syntax defined distance topologically where the distance between spaces
is the number of steps (spaces) that must be traversed (see the green routes in Figure 3.4).
For axial line (and to some degree natural street) analyses, these steps can be interpreted as
changes in direction; as (directly) onward travel is along the same space so no distance is
accrued whereas any non (directly) onward travel is to a different space so a step is made.
However, with axial and street segments, directly onward travel beyond an intersection is
to another space so a topological step is always accrued. In this case, it is essentially
counting the number of intersections. This difference in realisation is not trivial as it
bifurcates the empirical literature which has focused on the former (turns) and generally
found encouraging evidence (Bugmann and Coventry, 2005, 2008; Hutcheson and
Wedell, 2009; Jansen-Osmann and Berendt, 2002; Sadalla and Magel, 1980); though other
studies have found mixed (Jansen-Osmann and Wiedenbauer, 2004, 2006) and
unfavourable (Briggs, 1973; Herman, Norton and Klein, 1986) results. This can be
compared to the studies which consider and find a relationship between distance and the
number of intersections (Sadalla and Staplin, 1980b; for an exception see Nasar, 1983; see
also Sadalla and Staplin, 1980a).
Figure 3.4: Shortest paths between two locations according to different impedances and with the street and path network constructed from different units.

Shortest paths between the two locations using angular (red), topological (green) and metric (blue) impedances on A) Street segments, B) Named streets, C) Axial lines, D) Minimal fewest lines (where there are two equal topological shortest paths), E) Axial segments, F) Self-organised natural streets representations.
The validity of topological distance is also questioned because of other (relatively recent) cognitive wayfinding research. That is, because until the late 1990s the ‘long-dominant framework’ (Montello, 1998) of spatial microgenesis by Siegel and White (1975) - based on Piaget and colleagues’ theory of spatial ontogenesis (e.g. Piaget and Inhelder, 1956) - posited that spatial knowledge develops in distinct stages from topological to topographical\(^1\). Given this, and that spatial knowledge appears to develop over a year (Evans, 1980) but our primary routes are determined quickly and can be in as few as 5 or 6 route attempts (Rogers 1970, cited in Golledge, 1999), it seemed likely our routes are topologically determined. However, as identified by Montello (1998), the idea of longitudinal sequential stages is contrary to much of the research. For example, research (e.g. Brunyé and Taylor, 2008; Foo et al., 2005; Gärling et al., 1981; Herman, Blomquist and Klein, 1987; Holding and Holding, 1989; Ishikawa and Montello, 2006; Klatzky et al., 1990; Loomis et al., 1993; Montello and Pick Jr, 1993; Ruddle et al., 2011) shows that people can synthesise topographic information relatively instantaneously. This can also occur concurrently with or before acquiring topological knowledge; and that this knowledge is quantitatively refined. As such, models of human wayfinding should incorporate topographic detail.

Although largely independent to this research, and in fact partly proposed to overcome the *segment problem* (see below), space syntax introduced angular distance (Turner, 2000; Dalton, 2001; Turner, 2001). Here, the distance between two spaces is measured by the sum of the angular changes at intersections along the route (see also the red routes in Figure 3.4). As such, it broadly follows from topological and the same research except that it recognises that turnings of different magnitudes can be cognised (Sadalla and

---

\(^1\) Topographical is used to refer to non-topological representations (e.g. they incorporate angular and/or metric spatial relationships) to avoid confusion with terms that are used elsewhere, such as ‘metric’ and ‘Euclidean’, which also have other meanings.
Montello, 1989). That said, the extent to which turnings of similar magnitudes are discernible is not incorporated (Montello, 2007; see also Turner, 2004). The use of angular analyses however does remedy the segment problem. That is, for axial segment topological analyses, when a linear space (axial line) is broken into its constituent segments there is a step cost between each segment despite being linearly connected and no true turning being made (see also Frith, 2017). In contrast, using angular distance, linearly-connected axial segments cost nothing to transfer between as the deflection angle is 0° (see also Turner, 2001, 2007).

In contrast to the geometric impedances that ignore physical distances, distance can also be defined metrically. Although less commonly used and arguably undervalued (Montello, 2007) in space syntax analyses, the metric distance between two spaces is the physical length of the route between them (see also the blue routes in Figure 3.4). It is worth noting that metric distance does not combine well with axial representations due to their abstraction of geographic-Euclidean form. Also, and as noted earlier and despite its simplicity and relative plausibility, there is a volume of research indicating that people’s cognitive representations of distances are inconsistent with metric distance. That said, and while features such as turns (see earlier) may distort the relationship between actual and cognised metric distances and that the relationship may be linear (Cadwallader, 1973; Day, 1976; Howard, Chase and Rothman, 1973), non-linear (Briggs, 1973; Sherman, Croxton and Giovanatto, 1979) or either (Canter and Tagg, 1975; Wiest and Bell, 1985); they are often strongly correlated. For example, with correlation coefficients above 0.80 (MacEachren, 1980) and 0.90 (Cadwallader, 1973; Canter and Tagg, 1975; Howard, Chase and Rothman, 1973). Support for metric distance can also be found in the research cited as supporting turns-based distance as regardless of the number of turns, longer paths are
(still correctly) recalled as longer than the shorter paths (e.g. Hutcheson and Wedell, 2009; Jansen-Osmann and Wiedenbauer, 2006).

While there exist principal impedances for which each analysis assumes every individual follows the same logic, and they can in fact generate very similar routes (e.g. Raford, Chiaradia and Gil, 2005), it must be noted the literature suggests there may be systematic differences between individuals. For example, strangers may first navigate according to a topological or angular distance before developing metrically optimum paths as they become more familiar with an area (Zimring and Dalton, 2003; Turner, 2009b). This even extends to where it is suggested that complete newcomers to an area may navigate using highly local-connected vertices (degree) before adhering to the routes predicated by global measures such as closeness or betweenness (Haq and Zimring, 2003).

Also, and while these impedance attributes have so-far been discussed in terms of their use in dictating shortest paths, they are also used in determining radii. That is, as explained earlier, to model and include only realistic movement (lengths). Setting radii using metric distance can easily follow other research. For example, the mean walking journey length found in national travel surveys can be used for this purpose. However, for other measures of impedances, calibration is less straightforward. For example, how much angular change can be considered the limit to a normal pedestrian journey? That said, axial line analyses often set topological radii, typically where local equates to lines up to three steps away, but this is largely an arbitrary limit. Furthermore, using metric radii with non-metric impedances appears contradictory as it states that a person navigates using one logic but limits journeys according to another. This also causes other problems such as illogical destination and/or route choices (see Cooper, 2015).

Using radii can also remove ‘edge effects’ from an area of interest. That is, without radii, unless the study area is self-contained (e.g. an island), whatever boundary is used there
will be units outside it that are not included in the metric calculations. If these external units are not uniformly or symmetrically distributed, their omission will bias the results. However, by adding a buffer and including that area around the study area (that is of interest) for computation purposes only and then discarded, all units that can affect the study area are included in the metric calculations. This is not to say ‘edge effects’ no longer exist but they only effect the ‘buffer area’ which is subsequently removed. It is also worth noting that in non-radial analyses, the setting of the wider area around the study area (also called the catchment area) is a point of contention for the reasons explained above. Despite this, some studies (including where radii are used) analyse their study area as an independent area (without a catchment area) and therefore omit the influence of other units.
## Appendix A. 22 reviewed land use movement modelling studies

<table>
<thead>
<tr>
<th>Authors</th>
<th>Study Area</th>
<th>Predicted Variable</th>
<th>Predictor Variables</th>
<th>Main Preferred Model</th>
<th>Type</th>
<th>Data</th>
<th>Sites</th>
<th>Adj. $R^2$</th>
<th>Other Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pushkarev and Zupan (1975)</td>
<td>Manhattan, NY (US)</td>
<td>Pedestrians</td>
<td>Distance to nearest transit station (ft)</td>
<td>LR</td>
<td>LR</td>
<td>Avenue (midday)</td>
<td>344</td>
<td>0.35</td>
<td>St. error: 43.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Office floor space (ft$^2$)</td>
<td></td>
<td></td>
<td>Avenue (PM)</td>
<td>261</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Restaurant floor space (ft$^2$)</td>
<td></td>
<td></td>
<td>Street (midday)</td>
<td>228</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Retail floor space (ft$^2$)</td>
<td></td>
<td></td>
<td>Street (PM)</td>
<td>179</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Walkway space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behnam and Patel (1977)</td>
<td>Milwaukee, WI (US)</td>
<td>Pedestrians</td>
<td>Manufacturing Space (ft$^2$)</td>
<td>LR</td>
<td>LR</td>
<td>Hourly extrapolated from 6-minute counts</td>
<td>20</td>
<td>0.24</td>
<td>$R^2 = 0.60$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Parking Space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Residential Space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vacant Space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commercial space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cultural space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Entertainment space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Office space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Storage and Maintenance Space (ft$^2$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qin and Ivan (2001)</td>
<td>Connecticut (US)</td>
<td>Logged pedestrians</td>
<td>Has sidewalk</td>
<td>LLR</td>
<td>LLR</td>
<td>Non-inclement weather, weekday, and weekend count, between 8am-5:30pm</td>
<td>32</td>
<td>0.80</td>
<td>Also ran other models with more predictors: Model 1 had 13 (adj. $r^2 = 0.89$); Model 2 had 6 (adj. $r^2 = 0.82$) and Model 3 had 5 (adj. $r^2 = 0.81$). RMSE is also calculated which for the preferred mode is 0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Road is 2-lane highway</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Is campus area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Is tourist or downtown area</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pulugurtha and Repaka (2008)</td>
<td>Charlotte, NC (US)</td>
<td>Pedestrians</td>
<td>Total population within 1mi</td>
<td>LR</td>
<td>LR</td>
<td>7am-7pm</td>
<td>176</td>
<td>0.20</td>
<td>For 7am-7pm also ran similar models with 0.25mi (adj. $r^2 = 0.17$) and 0.5mi (adj. $r^2 = 0.13$). Models were also selected between based on Mallows $C_p$ (0.25mi = -4.0, 0.5mi = -</td>
</tr>
</tbody>
</table>
Chapter 3: Introduction to configurational methods

7.9, 1mi = -9.2 (Mallows 1973). Models were also run with same buffers for 7-8am, 10-11am, 12-1pm and 5-6pm with adj. $r^2$ of 0.13-0.19, 0.11-0.21, 0.13-0.21 and 0.08-0.13 respectively.

<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>Type</th>
<th>Variables</th>
<th>Model Type</th>
<th>Period</th>
<th>$N$</th>
<th>Adj. $r^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liu and Griswold (2009)</td>
<td>San Francisco, CA (US)</td>
<td>Pedestrians</td>
<td>Presence of bike lane, Job density within 0.25mi, Residential use within 0.0625mi (1/16mi), Transit stops within 0.375mi (3/8mi), Population density within 0.5mi, Mean slope within 0.0625mi, Diversity of land uses within 0.0625mi</td>
<td>LR</td>
<td>Weekdays 2:30-6:30pm</td>
<td>63</td>
<td>0.72</td>
</tr>
<tr>
<td>Schneider, Arnold, and Ragland (2009)</td>
<td>San Francisco, CA (US)</td>
<td>Pedestrians (crossing volumes)</td>
<td>No. rail transit stations within 0.1mi, Population within 0.5mi, Employment within 0.25mi, No. commercial properties within 0.25mi</td>
<td>LR</td>
<td>Extrapolated weekly volumes. From: Tue Wed Thu 12-2pm or 3-5pm; Sat:9-11am, 12-2pm or 3-5pm</td>
<td>50</td>
<td>0.90</td>
</tr>
<tr>
<td>Jones et al. (2010)</td>
<td>San Diego, CA (US)</td>
<td>Pedestrians</td>
<td>Employment density within 0.5mi, Population density within 0.25mi, Is there retail located within 0.5mi</td>
<td>LR</td>
<td>Weekdays, 7-9am</td>
<td>80</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Also estimated 11 other models, some with fixed buffers for all variables and some using backward elimination. Adj. $r^2$ ranged between 0.49-0.75.

Also ran other models including two using stepwise method (adj. $r^2 = 0.51-0.94$) and two others (adj. $r^2 = 0.46-0.47$). Also suggested "refinement factors" where predictions could be adjusted based on a weight where a criterion is met. Final model had a mean residual of -5.
<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>Model Type</th>
<th>Variables</th>
<th>Score Allocation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones et al. (2010)</td>
<td>San Diego, CA (US)</td>
<td>Pedestrian: Origin and Destination Model (Generators)</td>
<td>% Pedestrian Commuters, Person per residential acre, Employees per non-residential acre, Population over 65 per residential acre, Population under 16 per acre, Disabled population per residential acre, Mixed land uses</td>
<td>Weekdays, 7-9am</td>
</tr>
<tr>
<td>Jones et al. (2010)</td>
<td>San Diego, CA (US)</td>
<td>Pedestrian: Origin and Destination Model (Attractors)</td>
<td>Major transit center, Major transit stops, Other transit stops, Elementary schools, Universities and Colleges, Middle Schools, Neighbourhood Civic Facilities, Retail Facilities, Parks and Recreation, High Schools</td>
<td>Weekdays, 7-9am</td>
</tr>
<tr>
<td>Miranda-Moreno and Fernandes (2011)</td>
<td>Montreal (Canada)</td>
<td>Logged pedestrians (on intersection s)</td>
<td>% Major arterials within 400m, Bus stations within 150m, Commercial space within 50m (m²), Is four-way intersection, Logged distance to downtown, Open Space within 150m (m²), Population within 400m, Schools within 400m, Street segments within 400m, Subway stations within 150m, Very cold (min temp &lt;= 20°C), Very warm (max temp &gt; 32°C)</td>
<td>Aggregate (8hour). From: weekdays 6-8am, 11am-1pm, and 3:30-6:30pm</td>
</tr>
<tr>
<td>Study</td>
<td>Location</td>
<td>Pedestrians</td>
<td>Population within 400m</td>
<td>Commercial within 50m (m²)</td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>---------------------------</td>
<td>-------------</td>
<td>-------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>Miranda-Moreno, Morency, and El-Geneidy (2010; see also 2011)</td>
<td>Montreal (Canada)</td>
<td>Pedestrians (at intersection s)</td>
<td>509</td>
<td>0.54</td>
</tr>
<tr>
<td>Haynes and Andrzejewski (2010)</td>
<td>Santa Monica, CA (US)</td>
<td>Pedestrian</td>
<td>Employment density within 0.33mi</td>
<td>PM bus frequency</td>
</tr>
<tr>
<td>Lindsey et al. (2012)</td>
<td>Minneapolis, MN (US)</td>
<td>Pedestrians</td>
<td>% block residents that are non-white</td>
<td>% block residents over 65 or under 5</td>
</tr>
<tr>
<td>Kim, Ko, and Lee (2013)</td>
<td>Seoul (South Korea)</td>
<td>Pedestrians</td>
<td>Population within 300m</td>
<td>No employees within 300m</td>
</tr>
<tr>
<td>Study</td>
<td>Location</td>
<td>Target Group</td>
<td>Targets</td>
<td>Method</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------</td>
<td>--------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
</tr>
<tr>
<td>Schneider et al. (2013)</td>
<td>Alameda County, CA (US)</td>
<td>Logged Pedestrians</td>
<td>Households within 0.25mi, Employment within 0.25mi, Is in high-activity zone, Maximum slope of approaches, Within 0.25mi of university campus, Is controlled by a traffic signal</td>
<td>LLR (extrapolated from 2hr manual counts) weekly volumes. From: Tue, Wed or Thu 4-6pm or 7-9am</td>
</tr>
<tr>
<td>Tabeshian and Kattan (2014)</td>
<td>Calgary (Canada)</td>
<td>Pedestrians</td>
<td>No bus stops within 0.1mi, Total length of streets within 0.5mi, Total bus-km of routes within 0.75mi, No. dwellings within 0.5mi, Commercial space within 0.25mi (ha), No schools within 0.5mi, Total length of paths within 0.25mi</td>
<td>PR 7-9am, 11-1pm and 4-6pm in each year from 2007 to 2012</td>
</tr>
<tr>
<td>Mohamad et al. (1998)</td>
<td>Indiana (US)</td>
<td>Vehicles</td>
<td>Is locale urban, Is there easy access to state highway, Log of county population, Log of total arterial mileage of county</td>
<td>LR AADT</td>
</tr>
<tr>
<td>Xia et al. (1999)</td>
<td>Broward County, FL (US)</td>
<td>Vehicles</td>
<td>Are other county roads nearby, Number of lanes</td>
<td>LR AADT on county roads</td>
</tr>
</tbody>
</table>
### Chapter 3: Introduction to configurational methods

<table>
<thead>
<tr>
<th>Study</th>
<th>Location</th>
<th>Variables</th>
<th>Method</th>
<th>AADT</th>
<th>r²</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhao and Chung (2001)</td>
<td>Broward County, FL (US)</td>
<td>Vehicles, Area type (1=rural, 2=CBD or fringe, 3=residential, 4=outlying business district), Road functional classification, No of automobiles within a distance, No service employment within a distance</td>
<td>LR</td>
<td>816</td>
<td>0.76</td>
<td>Validation was performed on 40 extra data points (they excluded 4 outliers) showed the prediction differed from the observed on average by 23%.</td>
</tr>
<tr>
<td>Zhao and Park (2004)</td>
<td>Broward County, FL (US)</td>
<td>Vehicles, No. of lanes, Accessibility to regional employment, Direct access to expressway, Distance to regional centre of population, Population within variable-sized buffer, Employment within variable-sized buffer</td>
<td>GWR</td>
<td>816</td>
<td>0.87</td>
<td>Also ran an OLR to compare but this had lower adj. r² (0.76) and larger mean square errors (55.8 compared to 35.6). Validation also computed on 82 other data points with fewer and smaller errors for GWR than OLR.</td>
</tr>
<tr>
<td>Anderson, Sharfi, and Gholston (2006)</td>
<td>Anniston, AL (US)</td>
<td>Vehicles, Roadway functional classification, No. lanes, Population within 0.5mi, No employees within 0.5mi, ‘Through street’ or ‘destination street’</td>
<td>LR</td>
<td>58</td>
<td>0.80</td>
<td>38 other sites were used for validation with prediction r² values of 0.76. A paired t-test was also used which showed the difference between predicted and observed includes zero. They also compared to the current system (for 43 data points)</td>
</tr>
</tbody>
</table>
Yang, Wang, and Bao (2014) in Mecklenburg County, NC (US) vehicles:

- % residents in zip code below poverty line
- Car intensity: no. cars/road length (as below)
- Cars on the road (from satellite imagery)
- Median income
- No. housing units
- No. lanes

| LR | AADT | 243 | 0.65 |

Also reported AIC of -250.71; and ran their model without cars and car intensity (adj. $r^2 = 0.47$; and compared this model (from backward stepwise and SCAD shrinkage methods) to forward stepwise method. They also compare a variant to models in Zhao and Chung (2001) through validation (building models based on 200 sites to predict rest; repeated 1000 times) and found their model is better.
Appendix B.  78 reviewed analyses of movement and configurational methods

Campos, 1997
Can, 2012
Caria, Serdoura and Ferreira, 2003
Chang and Penn, 1998
Chiaradia, 2007
Cooper and Chiaradia, 2015
Cooper, 2015
Dai and Yu, 2014
Darjosanjoto and Brown, 1999
Dawson, 2003
Desyllas and Duxbury, 2001
Do et al., 2013
Eisenberg, 2007
Gil, 2012
Gil, 2014
Hillier and Iida, 2005
Hillier et al., 1987
Hillier et al., 1989
Hillier et al., 1990
Hillier et al., 1993
Hillier, Greene and Desyllas, 2000
Hillier, Yang and Turner, 2012
Hillier, 1993
Hillier, 2007
Hossain, 1999
Jiang and Jia, 2011
Jiang, Zhao and Yin, 2008
Jiang, 2006
Karimi and Motamed, 2003
Kasemsook, 2003
Kubat et al., 2012
Kubat, Atalay and Ozer, 2013
Kubat, 2001
Law, Sakr and Martínez, 2014
Lee and Seo, 2013
Lerman, Rofe and Omer, 2014
Liu and Jiang, 2012
Liu, 2007
Magalhães, 1997
Mahdzar, 2013
McCahil and Garrick, 2008
O’Neill et al., 2006
Oni, 2009
Ozbil and Peponis, 2007
Park, 2009
Parvin, Ye and Jia, 2007
Paul, 2012
Penn and Turner, 2001
Penn et al., 1998
Penn, 1993
Peponis et al., 1989
Peponis, Ross and Rashid, 1997
Perdikogianni and Penn, 2005
Preciado, 2012
Qiang, Miaoyi and Xingyi, 2015
Raford and Ragland, 2005
Raford, Chiaradia and Gil, 2007
Raford, 2003
Rashid, 1997
Read, 1999
Schwander and Law, 2012
Shetty, 2006
Song, 2013
Ståhle, Marcus and Karlström, 2005
Stonor, Campos and Smith, 2001
Tedjo and Funahashi, 1999
Teklenburg, Timmermans and Wagenberg, 1992
Trova et al., 1999
Turner and Dalton, 2005
Turner, 2005
Turner, 2007
Turner, 2009a
Varoudis et al., 2013
Wang, Rao and Feng, 2012
Wang, 2009
Wedderburn and Chiaradia, 2011
Ye, 1999
Zhang, Zhuang and Dai, 2012
Meta-analysis of movement flow estimations using configurational methods

The aim of this chapter is to provide an objective and empirical synthesis of the extent to which the configurational methods discussed in Chapter 3 correlate with pedestrian and vehicular movement flows. This includes synthesising all the main configurational methods (as described in Chapter 3) together to produce an overall estimate of the accuracy of these methods along with separate syntheses to highlight which particular configurational method (parameters) correlate with each type of movement. This work is motivated by the absence of such estimates in the literature so-far and the desire therefore to, for the first time, systematically identify which methods (parameters) best correlate with each type of movement. The contribution of the research reported in this chapter is two-fold. First, it will inform future work by finding the most suitable methods and parameters (i.e. those that best correlate with each type of movement) that should be used in studies that employ these methods. And, second, it will show which methods should be progressed in future research and/or require further investigation before they should (continue to) be used in research concerned with movement.
Chapter 4: Meta-analysis of movement flow estimations using configurational methods

4.1. Introduction

When applied to the estimation of pedestrian and vehicular movement flows, the configurational methods described in Chapter 3 are often reported as successfully explaining between 50-80% of the variance (Hillier and Vaughan, 2007; Penn, 2003; Penn and Turner, 2001). However, there are two key issues with existing evaluation of the methods.

The first issue is that the methodological and statistical rigour of the estimates is unclear. For example, from the limited information provided in Hillier and Vaughan (2007) and Penn and Turner (2001) where they refer to previous analyses, the estimates appear to be based on a narrative review. Whilst useful in some circumstances, alternative approaches such as meta-analyses offer significant advantages. For example, unlike narrative reviews, the process of deriving estimates and pooling them is transparent and lacks any significant subjectivity. The results are also empirically calculated so they have known statistical properties where, for example, the uncertainty of the estimates can also be computed.

The second issue is that the configurational analyses in question can be derived using a large number of parameters and combinations of parameters as shown in Table 4.1. However, beyond an acknowledgement that the parameters used can be influential (e.g. Penn, 2003) and a limited comparison of the influence of parameter selection in some studies (e.g. Hillier and Iida, 2005), it is unclear from the above estimates which parameters generally lead to better or worse correlations with pedestrian and vehicular movement.

The present study seeks to address these issues using a meta-analytic approach to statistically derive a pooled estimate of the accuracy of the various configurational analyses in predicting pedestrian and vehicular movement. Although the common approach for
Chapter 4: Meta-analysis of movement flow estimations using configurational methods

Table 4.1: Summaries of the key parameters in configurational analyses

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Types of spaces</strong></td>
<td></td>
</tr>
<tr>
<td>Street Segments</td>
<td>Sections of the road network between intersections and/or road end-points</td>
</tr>
<tr>
<td>Named Streets</td>
<td>Aggregations of street segments which are semantically continuous (share the same name)</td>
</tr>
<tr>
<td>Natural Streets</td>
<td>Aggregations of street segments which are angularly or visibly continuous</td>
</tr>
<tr>
<td>Axial Lines</td>
<td>Hand-drawn fewest and longest straight lines (of inter-accessibility) through the urban form</td>
</tr>
<tr>
<td>Axial Segments</td>
<td>Sections of axial lines which have been broken where they intersect other axial lines (with resulting short dangles removed)</td>
</tr>
<tr>
<td><strong>Types of impedance</strong></td>
<td></td>
</tr>
<tr>
<td>Topological</td>
<td>The number of steps or spaces that must be traversed along a route</td>
</tr>
<tr>
<td>Angular</td>
<td>The amount of angular change experienced at intersections (between spaces) along a route</td>
</tr>
<tr>
<td>Metric</td>
<td>The physical length of the route</td>
</tr>
<tr>
<td><strong>Accessibility measures</strong></td>
<td></td>
</tr>
<tr>
<td>Degree</td>
<td>The number of spaces incident to a space</td>
</tr>
<tr>
<td>Control</td>
<td>The sum of the reciprocals of degree values of spaces incident to a space</td>
</tr>
<tr>
<td>Closeness</td>
<td>The inverse of the mean shortest distances from a space to all other spaces</td>
</tr>
<tr>
<td>Integration</td>
<td>The normalised topological closeness score of a space</td>
</tr>
<tr>
<td>Normalised Angular Integration [NAIN]</td>
<td>The normalised angular closeness score of a space</td>
</tr>
<tr>
<td>Betweenness</td>
<td>The number of shortest paths between all spaces that pass through a space</td>
</tr>
<tr>
<td>Choice</td>
<td>The betweenness score of a space divided by its distance from all other spaces</td>
</tr>
</tbody>
</table>

this involves a systematic review of the literature to identify all relevant correlations, the resulting pool of correlations\(^2\) is problematic. Firstly, and in part due to the number of possible combinations of parameters in these analyses (see above), most studies only analyse (or report analyses for) a very small subset of the analyses. This is particularly problematic for a meta-analysis intended to compare the pooled estimates from different types of analyses as missing data can give misleading results solely due to the datasets where each analysis was applied (Mavridis et al., 2014). One solution to this is to omit analyses unless results are present (or can be calculated) for every permutation of analysis.

\(^2\) Although beyond the focus of this study, a systematic review was conducted in May 2016 and involved searching 14 search-engines using various combinations of keywords and hand-searching three other repositories. In total 5,334 papers were reviewed resulting in the identification of 145 relevant papers.
However, given the range of possible analyses, this would result in the exclusion of all datasets. For this reason, the approach taken here was to perform new configurational analyses for London (UK) which were then compared with counts of pedestrian and vehicular movement. London was selected due to the availability of pedestrian and vehicular movement count data for the city. In the meta-analysis that follows, all possible combinations of the parameters described in Table 4.1 were used to construct the estimates of movement and these were correlated with the empirical counts of movement.

4.2. Methodology

4.2.1. Configurational analyses

To begin, separate sets of graphs to model vehicular traffic (and so only included roads likely usable by vehicles) and pedestrian traffic (and so only included roads likely usable by pedestrians) were generated for each of the five main representations of road networks (see also Table 4.1). For the vehicular graphs, these were generated using the Ordnance Survey [OS] ‘integrated transport network’ road network dataset, while the pedestrian network representations were generated using this dataset plus the OS ‘urban paths’ pedestrian-only path network dataset. To generate the street segment graphs, roads were split at intersections and ‘split roads’ - where stretches of road between intersections are represented as multiple roads - were combined. For the named street graphs, street segments were joined based on adjacent segments sharing the same road name (including road name qualifiers). For the natural street graphs, the street segments were merged if their deflection angle is below 30° - based on Figueiredo and Amorim (2005) - to form the natural street graphs. The axial line graphs were hand-drawn following the space syntax methodology (for example see Al-Sayed et al., 2014) with the axial segment graphs formed by splitting the axial lines at intersections. As configurational analyses are susceptible to edge effects, where nodes closer to the edge of the study area are missing
other nodes (outside the study area) that would otherwise be included in the derivation of the measures, each graph covers the observation locations (see below) plus a buffer area to address this issue. The size of this buffer area was set to 7.5km such that it extends beyond the largest radii (5km) used in the measure calculations. Street segment and axial line-based maps of the study area are shown in Figure 4.1.

Each of the graphs were then analysed using the commonly used (and applicable) graph theoretic and space syntactic measures described in Table 4.1. These include the degree, control, closeness, integration, NAIN, betweenness and choice measures. Excluding the degree and control measures, which are not applicable with radii as they only include neighbouring units, these were calculated using 12 radii that are either commonly used in the literature or are a gradation of those commonly used (no radii; topological radii of 3, 5 and 7 steps; and metric radii of 500m, 1,000m, 1,500m, 2,000m, 2,500m, 3,000m, 4,000m

**Figure 4.1:** Maps of the street segment (left) and axial line (right) road and path networks for the study area including an 8km buffer (see also top inset) and an extract from the Isle of Dogs (bottom)
and 5,000m). Where applicable\(^3\) these measures were also calculated using three types of impedance (angular, metric and topological). This resulted in 622 parameter combinations. Existing software cannot calculate all 622 combinations (see also Appendix A), and consequently a new program, based on Brandes’ betweenness centrality algorithm (2001), was written in Python for this purpose.

### 4.2.2. Observation data

For the purposes of the meta-analysis, data which will be described in more detail below were obtained from six datasets (five of which were used in previous studies of configurational analysis) collected as part of various studies conducted in London (see also Figures 4.2 and 4.3 for maps of the observation locations). In total, 545 vehicular and 852 pedestrian count observations were available for analysis across the six datasets.

The first dataset, as originally used in Hillier et al (1993), consisted of 239 pedestrian count observations around 10 sub-areas of Kings Cross. The dataset as provided was likely missing 16 observations (based on the provided observation IDs) for unknown reasons whilst a further 12 (of the 239) were excluded as they could not be georeferenced for all road network representations. Moreover, in the original study, the authors collected data on ‘moving adults’ and ‘moving children’ but used only the data on adults in their analyses. In the current study, the average number of ‘moving people’ (‘moving adults’ and ‘moving children’ combined) are used. However, it should be noted that the two measures correlate almost perfectly ($r \approx 1.00, p < 0.01$).

---

\(^3\) Due to how they are calculated: integration can only be computed for topological analyses; NAIN can only be computed for angular analyses of axial segments; and degree and control are not applicable or computed with the different impedances as they only consider if a unit is neighbouring or not (see also Chapter 3).
The second dataset, also from Hillier et al (1993), as provided consisted of 50 pedestrian count observations from two areas of the City of London. The supplied dataset and the original analyses however excluded 12 observations from a third area because those locations were argued to be not comparable to the others in Hillier et al. (1993). Because this reasoning is argued here to be not sufficient for excluding data, these missing observations were obtained and included in what follows. However, two of these 12 and an additional three from the original 50 had to be excluded as they could not be georeferenced for all road network representations. Due to the addition of the third area and the availability of data for this area, the data used concerns the average number of moving people between 12-2pm rather than the average of three time-periods (10am-12pm, 12-2pm and 4-6pm) as used in the original analyses. However, the two data (12-2pm and the average of the three time-periods) correlate very strongly ($r = 0.98, p < 0.01$).

The third dataset, as originally used in Penn et al (1998), contained 321 pedestrian and 235 vehicular count observations in Barnsbury, Clerkenwell and Kensington; of which the latter is also sub-divided into South Kensington and Brompton Road. The data provided were missing 58 observations which are omitted for a range of reasons (Iida, 2006). Although some were simply missing because the data was not present (21 observations), others were missing non-randomly which may bias any analysis. For example, 35 observations were omitted because they were situated on streets or cul-de-sacs where normal movement is not expected (Iida, 2006). Beyond these issues, for these data a further nine pedestrian and two vehicular observations were excluded because they could not be georeferenced for all road network representations. These data concern the average number of ‘moving adults’ and ‘moving vehicles’ counted during the five time-periods (8-10am, 10am-12pm, 12-2pm, 2-4pm and 4-6pm).
Chapter 4: Meta-analysis of movement flow estimations using configurational methods

The fourth dataset, as used in Penn et al (1998; see also Chang and Penn, 1997, 1998), concerned 108 pedestrian and 67 vehicular count observations around South Bank. Of these, 10 pedestrian and three vehicular observations were excluded because they could not be georeferenced for all road network representations. The fifth dataset, also used in Chang and Penn (1998), concerned 146 pedestrian count observations around the Barbican estate. However, as a number of the observations were on overpasses (~70) or other minor paths which were not present in the Ordnance Survey data used to generate the road networks, and because the georeferencing was from low resolution stylized axial maps, 81 observations were excluded because they could not be georeferenced for all road network representations. The fourth and fifth datasets were collected in the same time-periods as the third dataset and concerned the average number of ‘moving adults’ and, for the former, the average number of ‘moving vehicles’.

The sixth dataset originated from the UK’s national network of road traffic manual counts and concerned 219 vehicular observations on major roads (motorways and ‘A’ class roads) and 33 observations on minor roads (‘B’ and ‘C’ class and other roads) in the area of central London, which was approximately bounded by the other datasets. Of these, three major road and one minor road observations were excluded because they could not be georeferenced for all road network representations. The data concerned the average number of moving vehicles counted from 7am to 7pm.

4.2.3. Correlation methodology

To correlate the pedestrian and vehicular observation data with the configurational analyses, the observation locations were georeferenced to their corresponding configurational units of analysis (e.g. street segments). For the data from the third-sixth datasets this was relatively simple as the data were collected using ‘gate counts’ where an observer stands at a fixed location and counts people or vehicles passing an imaginary
gate. As such, the gate counts were simply geo-referenced to that single location. However, data in the first and second datasets were collected non-statically using ‘moving observers’ who travelled at a constant pace along a route counting the number of people or vehicles they encountered. Each observation can therefore relate to multiple spatial units of analysis (e.g. several street segments). The methodological issue then is how to allocate the counts to discrete units of analysis (e.g. specific street segments). Given that some (unknown) proportion of each count would likely have been encountered along each spatial unit, the combination of units that make up the route is treated as the unit of analysis\(^4\). The equivalent metric scores are therefore calculated by the sum of each unit’s score weighted by the length of each unit.

The correlations (to be pooled) between movement and every combination of parameters were conducted using the entire data (all respective observations) and, for the pedestrian datasets, for each dataset and the sub-areas within them separately. This was possible and was plausible because the datasets cover relatively distinct geographic areas (see Figure 4.2). For the vehicular data and due some data (from the sixth dataset) overlapping with the study areas from the other datasets, the vehicular datasets were reorganised by geographic area (see Figure 4.3) and separate correlations computed\(^5\). This procedure was used rather than just observing the overall correlations using the entire data because pooling multiple correlations enables the degree of variability between correlations to be assessed (see below). The pseudo-study areas were also not unlike those used in the literature (and many have been used) and in practical applications (for examples, see Space Syntax Limited (2011)) and so the results are more likely to reflect the ability of these

\(^4\) Due to possible aggregation effects, the effect of removing this data was tested but it made little difference to the overall pattern of results and so is not discussed further.

\(^5\) Other groupings were also explored but these made little difference to the overall pattern of results and so are also not discussed further.
analyses to correlate with movement than if their ability was inferred to smaller areas solely from the overall London-wide correlation.

In terms of the correlations conducted, unlike previous studies which generally use the parametric Pearson’s $r$ correlation coefficient or linear regression, in the analyses that follow, the non-parametric Kendall Tau-a rank correlation coefficient ($\tau$), is used. There are a number of reasons for this. First, the associations computed tend to reflect a non-linear relationship and parametric statistics are sensitive to and can be distorted by this. As such, to be used meaningfully, their use relies upon appropriate transformations of the data prior to analysis. Given the large number of transformations that would be necessary (including just those transformations previously used in this literature), many more permutations of the data would need to be correlated and separately pooled. In contrast, non-parametric correlation statistics consider only the ranks of the observations and so ignore any non-linearity. As such, because the transformations do not change the relative ranks of the observations, they do not affect the coefficient and far fewer correlations

Figure 4.2: Maps of all pedestrian observation locations by dataset/geographic study-area (left) and for each set of study sub-areas (right)
need to be computed to produce a reliable analysis. Kendall’s $\tau$ is preferred because its statistical distribution approaches a normal one faster with smaller sample sizes and yields more reliable standard errors than other non-parametric measures such as Spearman’s index (Kendall and Gibbons, 1990). The $\tau$ coefficient is also relatively simple to interpret as it represents the probability that for any set of observations, the configurational analysis values are in the same order as the movement flow values. The coefficient values range between -1 (perfect negative concordance), 0 (perfect discordance) and 1 (perfect positive concordance). That said, because $\tau$ coefficients are smaller in magnitude compared to other indices such as Pearson’s $r$, Greiner’s relation, $r_g = \sin(\frac{\pi}{2} \tau)$, is used to calculate $r_g$ and $r_g^2$ which are approximately of equivalent magnitude to $r$ and $r^2$ values respectively (Kendall, 1949).

4.2.4. Meta-analysis methodology

The meta-analysis was completed in three stages for each of the pedestrian and vehicular sets of correlations. In the first stage, each set of correlations are quantitatively pooled together to summarise the overall mean correlation between configurational analyses and

Figure 4.3: Maps of the vehicular observation locations by the original dataset (left) and by the categorised geographic study-area (right) and study sub-area (right inset).
movement flows for all parameter combinations. In the second and third stages, the
correlations for each analysis parameter (e.g. each measure) and each combination of
parameters are pooled to indicate their average correlations with movement flows.

Two-tailed $Z$-tests are used to determine if each pooled correlation differs significantly
from zero. To determine if there are any statistically significant differences between
pooled correlations (in stages two and three), $Q$-tests are used if there are more than two
correlations to be compared (see below), and if significant, pairwise comparisons of the
mean correlations are completed using one-tailed $Z$-tests (Borenstein et al., 2011; Cooper,
Hedges and Valentine, 2009). In the second stage, the $Z$-test $p$-values are adjusted using
the conservative and well-known Bonferroni correction due to the risk of multiplicity and
erroneous inferences from the number of correlations which are to be compared.
Similarly, but because there are far more comparisons (622), in the third stage the $p$-values
are also adjusted but this time using the still powerful, but also relatively conservative,
Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

These syntheses can be computed using two modelling frameworks. The first, the fixed-
effects model, assumes observations (the observed correlations) are all estimates of the
same underlying parameter (the true population correlation) and any observed differences
are due to sampling error (within-study variance). It estimates the underlying correlation
as a weighted average of the observed correlations where the weights are equal to the
inverse of their variance. The alternative, the random-effects model, explicitly
acknowledges that there may not be a single underlying population parameter and so
models any heterogeneity (or inconsistency) between observations. To do this, it
incorporates a between-study variance component, along with the within-study
component, and treats the observed correlations as if they are from a normally-distributed
population, estimating the mean and standard deviation (\(\text{tau}^6\)) of that population. Given this, the likelihood of a non-constant population parameter (Field, 2003) and that the fixed-effects model can be considered a special case of the random-effects model when between-study variance is low, this meta-analysis is completed using the random-effects models.

That said, heterogeneity between the observed correlations is investigated using three metrics. The first, the Cochran’s \(\chi^2\) \(Q\)-test, assesses whether the observed differences in the correlations are compatible with chance. This test is however sensitive to the number of correlations pooled and can fail to reject the null hypothesis of homogeneity when there are few correlations pooled. It can also detect trivial heterogeneity when there are many correlations. As such, following common practice \(\alpha = 0.05\) except for the third stage of this meta-analyses where \(\alpha = 0.10\) due to the low number of correlations pooled (23 and 12 for the pedestrian and vehicular meta-analyses respectively). The second metric, the \(I^2\) index, measures the percentage of the variation in the correlation estimates that is due to heterogeneity rather than sampling error. Values of around 25%, 50% and 75% suggest low, moderate and high amounts of heterogeneity, respectively (Higgins and Thompson, 2002). The \(I^2\) index however can be imprecise so is interpreted using its (95%) confidence intervals (von Hippel, 2015). Also, when the correlations pooled are based on large sample sizes and so have smaller sampling errors, the \(I^2\) index can be high even where the amount of heterogeneity is trivial (Borenstein et al., 2008). As such, heterogeneity is also assessed using a third metric, the tau-derived 95% prediction intervals, which provide an estimate of the range within which 95% of future similar correlations would be expected to lie. When large amounts of heterogeneity are identified,

---

6 To avoid confusion, Kendall tau correlation coefficients are presented in Greek script (\(\tau\)) and the tau measure of heterogeneity are presented in Latin script (\(\text{tau}\)).
the results should be interpreted cautiously. For one, because the pooled estimates simply
describe the population mean and not the spread of that population. Additionally, large
amounts of heterogeneity may also suggest a mixture of population distributions that
require subgroup analyses (Borenstein et al., 2011). This is particularly the case when there
are *a priori* known subgroups such as correlations with different parameters and different
combinations of parameters (in the first and second stages of this meta-analysis).

Lastly, in their standard form, random-effects models assume independence between
correlations which is likely violated in two ways in this meta-analysis. The first, called
hierarchical dependency, arises due to the nesting of study areas within study areas. This
occurs where it is not unlikely that correlations within the same nest are more similar than
those in different nests. Where this occurs, their correlations cannot be considered
independent. This problem is mitigated here by using multi-level models with the sub-
areas (where present) as the base level, study areas as the next level and the overall London
area as the apex level. The second violation, called sampling or statistical dependency,
arises because some correlations are based on the same or similar samples. This introduces
dependency at the sampling level where the sampling errors are correlated. In what
follows, this is accounted for by using multivariate models and specifying the Kendall’s ϱ
variance-covariance matrices (see also Cliff and Charlin, 1991; Gleser and Olkin, 2009)
following the generalised least squares estimation methods by Becker (1992; see also
Olkin, 1976). All meta-analyses are computed using the ‘metafor’ package (Viechtbauer,
2010) with the between-study variance estimated using the recommended restricted
maximum likelihood method (e.g. Thompson and Sharp, 1999; Viechtbauer, 2005) in the
statistical software environment R (R Core Team, 2016).
4.3. Results

The Forest plot in Figure 4.4 shows the estimates for the pedestrian and vehicular correlations for the overall pooled estimates (labelled overall) and pooled by each parameter (labelled by the parameter). As such, the overall pooled estimates (top row) show the overall mean correlation coefficient for all permutations of parameters along with the variation observed across parameter combinations. To investigate further, the forest plots shown in Figures 4.5 (for the pedestrian correlations) and 4.6 (for the vehicular correlations) show the strength of the pooled estimates of the five most strongly correlated parameter combinations and ten other analyses commonly used in the literature. Plotted in the figures are the pooled estimates of the mean $\tau$ coefficients (Figure 4.4) or the strength of the pooled mean coefficients (Figures 4.5 and 4.6) and their 95% confidence (the undashed line) and prediction intervals (the dashed line). Summary data in the figures note the number of correlations pooled ($N$); the pooled mean correlation coefficient and its $p$-value, the standard error and confidence intervals; the Tau and prediction intervals, the significance of the $Q$-test and the $I^2$ index and its confidence intervals. Figures 4.5 and 4.6 also show the rank position of its correlation strength (out of the 622 computed) where those ranked 1st are the strongest average correlating. The results for the pedestrian and vehicular meta-analyses are discussed in turn.

4.3.1. Pedestrian results

As shown in Figure 4.4, the pooled mean $\tau$ correlation between all configurational models and pedestrian traffic is 0.11 and this is significantly different to zero ($Z = 6.11, p <$

---

7 By removing the sign from the pooled coefficient the results are easier to interpret and the post-hoc comparisons identify differences in the mean strengths of the relationships (regardless of direction) rather than differences in the mean coefficients.
Figure 4.4: Forest plot showing the pooled correlation estimates of the overall and by parameter correlations with pedestrian (top) and vehicular (bottom) traffic.
0.01). This is approximately equivalent to a \( r \) correlation coefficient \((r_g)\) of 0.17. The significant \( Q \)-test \((p < 0.01)\), \( I^2 \) value of 100% (including confidence intervals) and relatively wide prediction intervals of -0.13 to 0.35 (which are approximately equivalent to \( r \) coefficients of -0.20 to 0.52) indicates there is substantial variation across the pooled correlations.

Considering the types of parameters first, there was a significant difference in the pooled correlations across the different types of units used to create the networks \((Q = 24.28, df = 4, p < 0.01)\). On average, the correlations that for the named street representation were significantly greater than those calculated using natural street \((Z = 3.08, p < 0.05)\), street segment \((Z = 3.16, p < 0.01)\) or axial line \((Z = 4.74, p < 0.01)\) representations. The mean correlation using axial lines was also significantly smaller than those using axial segments \((Z = 2.74, p < 0.05)\).

There was also a significant difference between the mean correlations across the accessibility measures \((Q = 136.22, df = 6, p < 0.01)\). The mean correlations for analyses using betweenness and NAIN measures, which had the largest pooled mean correlations, were significantly greater than those using choice, degree, control or closeness measures\(^8\). The mean correlation for the analyses that used closeness were also significantly smaller than those using integration \((Z = 4.04, p < 0.01)\) and choice \((Z = 4.30, p < 0.01)\).

---

\(^8\) The mean correlations of analyses using betweenness was significantly greater than those using choice \((Z = 3.32, p < 0.01)\), degree \((Z = 3.13, p < 0.05)\), control \((Z = 4.09, p < 0.01)\) and closeness \((Z = 7.73, p < 0.01)\) and likewise with those using the NAIN measure compared to those using choice \((Z = 3.58, p < 0.01)\), degree \((Z = 3.24, p < 0.05)\), control \((Z = 4.32, p < 0.01)\) and closeness \((Z = 9.23, p < 0.01)\).
There was also a significant difference between the mean correlations based on the impedance used ($Q = 7.16, \text{df} = 2, p < 0.05$). Those that used topological impedance were significantly greater than those using angular ($Z = 2.46, p < 0.05$) or metric ($Z = 2.54, p < 0.05$) impedances.

Finally, a $Q$-test also identified a significant difference in the mean correlations based on the radii used ($Q = 23.61, \text{df} = 11, p < 0.05$). However, pairwise $Z$-tests only identified that analyses calculated using a 2000m radii had on average significantly greater correlations than those using 1500m radii ($Z = 3.33, p < 0.05$).

Although these results indicate significant differences between the mean correlations, the results must be interpreted cautiously as there is evidence of large amounts of heterogeneity in the correlations pooled together and the estimates have relatively wide and overlapping prediction intervals.

In terms of the specific models or permutations of parameters, Figure 4.5 shows that the mean strength of the (three) best fitting analyses is 0.49. Of which, two have positive mean $\tau$ coefficients of 0.49 and the third has a negative mean tau of -0.49. These are approximately equivalent to $r$ values of 0.69 and -0.69 respectively. These three analyses, and the two next best fitting analyses (also shown in Figure 4.5), were all calculated using axial lines, angular impedance, various radii, and the betweenness or, in one model, the closeness measures. The strength of the correlation for the ten selected analyses$^9$ ranged between 0.38 and 0.47 but a $Q$-test found no significant difference between the mean

---

$^9$ These analyses (labelled by their rank position) are included as they are the strongest correlating analyses using: the named street (8th), natural street (28th), street segment (36th) and axial segment (92nd) representations; the choice (17th), integration (150th), degree (244th), NAIN (261st) measures; and using metric (8th) and topological (9th) impedance. The analyses ranked 205th and 261st are also included because they are the strongest correlating analyses closely related to those typically used in the (space syntax) literature.
correlation strengths of all the analyses shown in Figure 4.5 ($Q = 12.96, df = 14, p > 0.05$). Going further, a comparison of the average correlation strength from the best fitting analysis against all other analyses found only 380 analyses (61%) had a significant weaker average correlation strength. Significant $Q$-tests of homogeneity and large $I^2$ values (the smallest of which is 89% with confidence intervals of 84%-92%) for every pooled estimate indicates heterogeneous correlation strengths. That said, the prediction intervals vary in size for each analysis and suggests for some, for example the analysis ranked 1st-5th as shown in Figure 4.5 which have moderately-sized prediction intervals, the heterogeneity may not be critical.

### 4.3.2. Vehicular results

As shown in Figure 4.4, the pooled mean $\tau$ correlation between all configurational analyses and vehicular traffic was 0.09 and this was also significantly different to zero ($Z = 4.24, p < 0.01$). This is approximately equivalent to a $r$ coefficient ($r_g$) of 0.14. The
significant $Q$-test ($p<0.01$), very high $I^2$ value (including confidence intervals) and relatively wide prediction intervals of -0.12 to 0.30 (which are approximately equivalent to $r_g$ values of -0.19 to 0.45) indicates there is substantial variation in the pooled correlations.

Considering the types of model parameters, in terms of the units used to create the networks, there was no significant difference between the mean correlations of the five different units ($Q = 0.67, df = 4, p > 0.10$) which ranged between 0.06 and 0.10. The average correlation of analyses computed using natural streets was however not significantly different to zero ($Z = 1.12, p > 0.05$).

There was a significant difference ($Q = 125.88, df = 6, p < 0.01$) in the mean correlations based on the measure used. The NAIN measure, which had the largest average correlation, was significantly greater than all others except betweenness and choice$^{10}$. Betweenness which had the second largest mean correlation was significantly greater than the control ($Z = 3.40, p < 0.01$) and closeness ($Z = 6.13, p < 0.01$) measures. Also, closeness which had the lowest mean correlation was significantly lower than all other measures$^{11}$.

There was a significant difference ($Q = 17.75, df = 2, p < 0.01$) in the average correlations of the three types of impedances. Specifically, the mean correlations for the analyses computed using angular impedance were significantly greater than those

---

$^{10}$ The mean correlations of models using NAIN was significantly greater than those using closeness ($Z = 8.69, p < 0.01$), control ($Z = 4.85, p < 0.01$), degree ($Z = 3.80, p < 0.01$) and integration ($Z = 3.17, p < 0.05$).

$^{11}$ The mean correlations of models using closeness was significantly lower than those using betweenness ($Z = 6.13, p < 0.01$), choice ($Z = 3.51, p < 0.01$), control ($Z = 4.14, p < 0.01$), degree ($Z = 3.64, p < 0.01$), integration ($Z = 4.73, p < 0.01$) and NAIN ($Z = 8.69, p < 0.01$).
computed with metric ($Z = 4.20, p < 0.01$) or topological ($Z = 3.21, p < 0.01$) impedance.

There was also a significant difference ($Q = 78.10, df = 11, p < 0.01$) between the mean correlations calculated using the different radii. The pooled means of analyses computed with no radii and with 5000m and 4000m radii were significantly greater than analyses using the three topological radii (3 steps, 5 steps and 7 steps) and the 500m radii\textsuperscript{12}. The mean correlation of the 3000m radii was also significantly greater ($Z = 3.30, p < 0.05$) than that of 3 topological steps radii.

As with the pedestrian model parameter meta-analyses, these results however only indicate that there exist significant differences between the average correlations. That is there is evidence of large amounts of heterogeneity in the correlations pooled within each parameter. This includes, to some degree, the NAIN measure which had moderate-to-high amounts of heterogeneity according to its $I^2$ value of 71% and confidence intervals of 66% to 76%.

In terms of the specific models or permutations of parameters, Figure 4.6 shows that the mean strength of the (two) best fitting models is 0.54. This is approximately equivalent to a $r$ value of 0.75. These two models (and the next three best fitting models) were derived using the choice measure. In the best model, the choice measure is used with the natural

\textsuperscript{12} The mean correlations of models using no radii was statistically significantly greater than those using 3 topological steps ($Z = 5.99, p < 0.01$), 5 topological steps ($Z = 5.30, p < 0.01$), 7 topological steps ($Z = 4.37, p < 0.01$) and 500m radii ($Z = 4.46, p < 0.01$) and likewise with those using 5000m radii compared to those using 3 topological steps ($Z = 4.89, p < 0.01$), 5 topological steps ($Z = 4.33, p < 0.01$), 7 topological steps ($Z = 3.57, p < 0.05$) and 500m radii ($Z = 3.60, p < 0.05$) and those using 400m radii compared to those using 3 topological steps ($Z = 4.31, p < 0.01$), 5 topological steps ($Z = 3.84, p < 0.01$), 7 topological steps ($Z = 3.21, p < 0.05$) and 500m radii ($Z = 3.20, p < 0.05$).
street representation, angular impedance and no radii while the second best fitting model was computed using the named street representation, metric impedance and a 5000m radii. The majority of the other best fitting models also tended to be calculated using natural or named streets, and in some cases street segments, using the choice, and in some cases the betweenness, measures. They were also often calculated using no radii or a 5000m radii; although there tend to be no significant differences between the equivalent models with these two radii. For example, the models ranked 2nd and 5th did not differ \((Z = 0.10, p > 0.05)\) and the model ranked 1st and its equivalent with a 5000m radii \((Z = 0.30, p > 0.05)\) did not differ.

The strengths of the other best fitting models (in the top 5) are all 0.53 and the other selected models\(^{13}\) as shown in Figure 4.6 have mean correlation strengths of 0.30 to 0.52.

A \(Q\)-test of the mean correlation strengths of these 15 models indicates there is a significant difference in the pooled means \((Q = 44.07, df = 14, p < 0.01)\) and

Figure 4.6: Forest plots showing the modulus of the pooled correlation strength estimates between selected specific analyses (permutations of parameters) and vehicular traffic

![Forest plots showing the modulus of the pooled correlation strength estimates between selected specific analyses (permutations of parameters) and vehicular traffic.](image)

Note: The horizontal axis represents correlation strength and so the plotted results ignore the direction of the relationship and truncates the results to 0,1. The figures in the included table are the raw pooled \(\tau\) coefficients.

\(^{13}\) (space syntax) literature.
pairwise Z-tests identified that the mean strength of the model ranked 345th (that uses topological integration analyses of axial lines with a topological radii of 7 steps) is significantly lower than that of the models ranked 1st \((Z = 3.68, p < 0.05)\), 2nd \((Z = 3.36, p < 0.05)\), 3rd \((Z = 3.84, p < 0.01)\), 4th \((Z = 3.70, p < 0.05)\), 5th \((Z = 3.36, p < 0.05)\) and 7th \((Z = 3.34, p < 0.05)\). Furthermore, comparisons of the pooled average correlation strength of the best fitting model to all other models indicates that only 435 models out of the 621 (70\%) had significantly weaker correlation strengths.

According to Q-tests for which 98\% of the tests were significant and \(I^2\) values were 82\% or above (and in most cases this includes the lower bound of its confidence interval), the vast majority of models exhibited significant heterogeneity between their correlations. That said, the widths of some prediction intervals were low, for example 0.01 in the case of the model ranked 3rd (in Figure 4.6), and this may indicate for some models the heterogeneity is not meaningful.

### 4.4. Discussion

The aim of this chapter was to explore the ability of various configurational models to correlate with pedestrian and vehicular movement flows. To do this, various movement count data were taken and correlated against each of the configurational models. These correlations were then synthesised in a random-effects meta-analysis to examine the average correlations of these models along with the variance of the correlations in different datasets.

This study found the average \(\tau\) correlations between (all) configurational models and pedestrian and vehicular movement were 0.11 and 0.09 respectively which approximate to very weak but significant \(r\) coefficients of 0.17 (0.14) and very small \(r_g^2\) values of 0.03 (0.02). Although caution must be used when using the approximated \(r_g^2\) metrics, they are
highly inconsistent with those from the literature which suggest $r^2$ values between 0.50-0.60 to 0.80 for both types of movement (Hillier and Vaughan, 2007; Penn, 2003; Penn and Turner, 2001). That being said, it is likely these previous studies were intended to present an approximate range of ‘best case’ results rather than the ‘average result’. Previous findings therefore likely only relate to only specific parameters or combinations of them, whereas all possible permutations were explored here. Furthermore, the small mean correlations found in this meta-analysis may be attributable to the use of a random-effects model which assumes that all correlations, regardless of model parameters, belong to the same population, which may not be true. For example, the observed heterogeneity between correlations in this estimate can be explained because some parameters tend to yield negative correlations (e.g. closeness analyses) and some combinations of model parameters (e.g. topological analyses of street segments) yield negligible correlations. In both cases, even if the majority of the correlations and the general relationship is positive, these correlations will bias the mean correlation towards 0.

A better evaluation of the configurational analyses is to consider the correlations with each of the 622 combinations of parameters as distinct populations of effects. In this way and taking the combinations of model parameters which best correlate with pedestrian and vehicular movement, the findings presented in this study suggest that these models have average $\tau$ coefficients up to 0.49 (and -0.49) and 0.54 respectively. When approximated as Pearson’s $r$ coefficients, these can be interpreted as strongly correlating with $r_g$ values of 0.69 (and -0.69) and 0.75 and $r_g^2$ values of 0.48 and 0.56 respectively. Although not entirely dissimilar to the estimates from the previous studies (see above), the confidence intervals of the top fitting pedestrian (0.41-0.56) and vehicular (0.48-0.59) models only marginally overlap with the range, 0.50-0.60 to 0.80, estimated in previous reviews (Hillier and Vaughan, 2007; Penn, 2003; Penn and Turner, 2001). This suggests
that the results presented in this analysis meaningfully differ to the previous estimates and that the previous estimates were likely somewhat optimistic.

Another finding from this meta-analysis is the non-significant difference between the correlation strengths of the models that best correlate with pedestrian and vehicular traffic and around 40% and 30% of the respective next best correlating models. That is, there is an insufficient difference, relative to the sampling errors, to statistically discern between the best correlating model and 241 (pedestrian) and 186 (vehicular) other models. That being said, there are some parameters and combinations of parameters that either tend to better correlate and/or similarly correlate and have other advantages. Firstly, whilst combinations of axial lines (and axial segments) and other parameters yield some of the top correlating models (particularly for pedestrian movement), they do not significantly improve over the non-axial alternatives. This is an important because the generation of networks for axial analyses of a non-trivial urban area is labour intensive and not inherently reproducible. Taking this into account, the current findings suggest that purely for the purposes of estimating movement, the axial methodology is perhaps unnecessary.

Secondly, the top combinations of parameters for predicting pedestrian and vehicular movement tended to employ the betweenness and (its normalised equivalent) choice measures. They also tended to define distance angularly (or metric in some cases). Lastly, whilst for the pedestrian analyses a range of radii centred around 2km-3km resulted in the best correlations; for the vehicular analyses, the top radii were consistently either none or 5km. In fact, although vehicular analyses with no radii tended to produce marginally better correlating models, there was often no significant difference to those using a 5km radii. As such and given models with larger or no (in that all units are included in the metric calculations) radii require greater computation time (due to the larger number of nodes
involved in the measure calculations which would otherwise be excluded as they are beyond the radii); models using the 5km radii are arguably more practical.

Beyond these point estimates of the mean correlations of different models, there was some evidence of heterogeneity between the correlations. This was also present when pooling (vehicular or pedestrian) models with the same combination of parameters where it may have been assumed they would possess relatively similar (homogeneous) correlations. The consequence of this is that although the models (as discussed) may, on average, yield higher (or lower) correlations than other models, it does not guarantee the same for a future correlation (or correlations). The extent of this can be judged by tau and the 95% prediction intervals. In the case of the pedestrian models, tau and the prediction intervals for the top five models are all relatively similar. For the vehicular models, for example, the model ranked 3rd (angular choice analyses of street segments with no radii) has a considerably smaller tau and prediction intervals than those ranked 2nd and 5th. This indicates the strength of a future correlation using the model ranked 3rd is likely to be between 0.49 and 0.58 whereas for the models ranked 2nd and 5th it is 0.29-0.78 and 0.32-0.74. In light of this, although there is a probability that the models ranked 2nd and 5th will yield stronger correlations than the model ranked 3rd, there is also a probability it will be less strong. As such, and although the importance of this attribute may vary for different purposes, the vehicular models ranked 3rd, and to some degree 4th and 1st, ought to be preferred as they assure a greater minimum and consistent correlation strength for a new dataset.

It is important to reiterate and acknowledge how the approach taken in this analysis differs from that commonly used in the literature and its general limitations. The first major difference is the use of the Kendall τ rank correlation coefficient rather than Pearson’s $r$ or linear regression. That is, and beyond that discussed earlier, while a similar meta-
analysis could pool $r$ and $r^2$ values to analyse the average strength of the relationship; for the purposes of using these methods to forecast movement flows for a new study area, they are somewhat meaningless unless the associated regression coefficient is relatively consistent or can be determined \textit{a priori}. Neither of these have been shown to be true in the literature and so in the absence of this, the pooled $r$ and $r^2$ values are only applicable if (some) movement flow observations are used to train the coefficient for that particular study area. Also related to this, pooling the results from the regression of movement flows using network measures is likely to be upwardly biased. This is because the model parameters are trained to the current dataset, including its random noise, and so only describe the extent to which they fit that particular sample. As such, they should not be used as indicators of how well they would fit new samples, or in the case of a meta-analysis, pooled to indicate the population-wide fit. It is in this situation, and (all) analyses in general, that cross-validation should be used whereby the models are trained on one portion of the data and tested against another.

The results from this meta-analysis should also be considered in light of the data used and its generalisability. That is, despite the fact that the statistical approach taken to pool analyses is far more valid than relying on the results of a single analysis, all of the analyses relate to areas of London and particularly central London. As such, the results may have limited transferability to other areas in London or beyond. Moreover, and although the cherry-picking of data in space syntax analyses (see also earlier) was offset by the inclusion of vehicular count observations from non-space syntax sources in this analysis, the effects and remaining consequences of this are unknown at this stage. For example, are the included types of locations where count observations are available representative of London? Are they representative and the results transferable to elsewhere in the UK, the world or more generally to other societies and cultures? Future analyses may want to
consider this; and although this could be through the correlations identified in the systematic review of the literature, the aforementioned problems (see also earlier) would need to be considered and mitigated.

Having said this, this meta-analysis is, as far as the author is aware, the first of its kind in this literature. The principal finding is that these configurational models can approximately explain as much as 50-60% of the variance in pedestrian and vehicular movement flows and this figure is more conservative than that noted in previous studies. The best models for predicting these flows also differ from those typically used in related analyses. For example, angular betweenness analyses of the axial line representation of the road network and a 2000-3000m radii appeared to best correlate with pedestrian movement flows whereas angular choice analyses using natural streets and no radii did for vehicular movement flows. That said, there was very little difference in the correlations between many of the models and so the choice of configurational model, ignoring those that performed poorly, can be determined and justified on other factors. For example, betweenness analysis of street segments with a 1000-2000m radii (for pedestrian movement) and 5000m radii (for vehicular movement) could be equally effective and justified if a finer spatial granularity is required – as is the case for the analyses using these metrics in Chapters 10 and 11.
## Appendix A. List of computed configurational analyses

<table>
<thead>
<tr>
<th>Unit</th>
<th>Radii</th>
<th>Angular impedance</th>
<th>Metric impedance</th>
<th>Topological impedance</th>
<th>N/A</th>
<th>Control</th>
<th>Degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Axial Line</td>
<td>3</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>1000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>1500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>2000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>2500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>3000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>4000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>5000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>No radii</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td>Axial Segme</td>
<td>3</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td>nt</td>
<td>5</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>1000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>1500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>2000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>2500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>3000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>4000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>5000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>No radii</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td>Named Street</td>
<td>3</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>1000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>1500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>2000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>2500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>3000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>4000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>5000m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>500m</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>No radii</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td></td>
<td>N/A</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td>✓✓✓✓</td>
<td></td>
<td></td>
<td>✓✓</td>
</tr>
<tr>
<td>N/A</td>
<td>500m</td>
<td>5000m</td>
<td>3000m</td>
<td>2000m</td>
<td>1500m</td>
<td>1000m</td>
<td>500m</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Natural Street</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Street Segment</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
New configurational measures of movement

From the results presented in Chapter 4 and previous research, it is clear that configurational measures of road networks correlate with overall (pedestrian and vehicular) movement flows. These measures therefore have the potential to be utilised as proxy measures for the levels of potential guardianship provided by passers-by (see also 0). In this chapter, three novel sets of configurational measures are created to capture constructs in the criminological literature. These three sets of measures are proposed to:

1) account for the non-uniform distribution of trip generators around the network to provide more accurate estimates of movement flows; 2) estimate guardianship intensity rather than just the likely numbers of (potential) guardians; and 3) estimates of offender awareness spaces. These measures are then illustrated using a simple road network.

5.1. Introduction

It has been shown that configurational measures, and especially those employing the betweenness measure (see Chapter 4), can be used as estimates of pedestrian and vehicular movement flows around a road network (see Chapter 4). As suggested in earlier criminological research (e.g. Hillier, 2004; Davies and Johnson, 2015; Frith et al, 2017),
these metrics can be used to estimate the potential guardianship provided passively by passers-by. These methods, however, may be adapted to produce more nuanced estimates of movement and guardianship potential. In this chapter, three novel sets of betweenness-based graph theory metrics are proposed that:

1. Account for the non-uniform distribution of trip generators to provide more accurate estimates of movement flows.
2. Account for the capability of potential guardians to estimate qualitatively different levels of guardianship rather than the simple number of (potential) guardians.
3. Account for an offender’s likely asymmetric travel patterns to estimate their awareness space.

In the sections that follow, the rationale for each of these metrics will be discussed.

### 5.2. Modelling non-uniform distributions of trip generators

#### 5.2.1. Non-uniform distributions of trip generators

One limitation with traditional graph theory and space syntax measures (see also Chapters 3 and 4), including betweenness, is that they assume all vertices contribute equally in the metric calculations. In the case of betweenness which measures the degree of overlap of the shortest routes through the network, when used to predict movement, the values can be likened to measuring the overlap of peoples’ journeys throughout the network. As such, under the belief that all vertices contribute equally to the metric, it assumes all vertices (within a distance if a radius is used) are equally likely to be origins and destinations for journeys.

The problem is that some vertices, such as road segments with more or fewer properties, will likely act as origins and destinations for more or fewer journeys (e.g. Leung et al., 2011). These vertices will also be non-uniformly distributed through the network. For
example, vertices with more properties will tend to be found in certain areas such as city
centres and other built-up areas. In contrast, those with fewer properties will tend to be
found on the out-skirts and in more rural areas. Ignoring this can result in under- (over-)
predicting the number of trips to and from busy (rarely used) road segments. In effect,
the degree of overlap of journeys through the network and therefore the estimated
amount of movement may be mis-estimated.

5.2.2. Vertex-weighted betweenness

One solution to this, based on that used elsewhere (e.g. Chiaradia, 2007; Turner, 2007), is
to incorporate vertex weights in the metric calculations. To explain, recall that betweenness
is typically calculated by (see also Equation 4 in Chapter 3):

\[ B_e = \sum_{i,j \in V: i \sim j} \frac{\sigma_{ij}(e)}{\sigma_{ij}} \]  (5)

Where the betweenness score for each vertex \( e \) is the amount of overlap in the shortest
routes between all vertices \( (i \text{ and } j) \) that pass through it (see also Chapter 3). Because
some vertices act as origins or destinations for more journeys, a weight - which in previous
analyses (e.g. Chiaradia, 2007; Turner, 2007) are the lengths of the road segments
represented by the vertices – is applied to each vertex. The routes are then weighted by
the origin and destination vertex’s weights. This is because longer segments are expected
to generate more trips than shorter segments and so if the origin or destination segments
are shorter (or longer) then the vertices on the paths between those vertices will receive
less (more) weight. Given the origins’ \( w_i \) and destinations’ \( w_j \) weights, this equates to:

\[ W'_e = \sum_{i,j \in V: i \sim j} \frac{\sigma_{ij}(e) w_i w_j}{\sigma_{ij}} \]  (6)
However, an issue with this is that in the context of modelling movement when a radius is used, each origin may be capable of reaching a different number of destinations (a different total length of road). As such, and based on Equation 6, if a given origin vertex can reach more road, the formula is effectively assuming more trips are made from that vertex (regardless of its length) than from a vertex that can reach less total length of road. In reality though, it would be expected that if more road (in terms of being destinations) can be reached, there are not more trips made. Instead, that the trips are shared and so fewer trips are made per reachable destination (road length). To account for this, Equation 6 can be modified so that the weight applied to each vertex on routes through the network is divided by the total weights reachable from the origin ($\sum_{k \in V} w_k$). This means that if more destinations can be reached from an origin (compared to another origin) then a smaller weight is applied on each route to each destination. This vertex-weighted betweenness measure would therefore be calculated by:

$$W_e = \sum_{i,j \in V, i \neq j} \frac{\sigma_{ij}(e) w_i \ w_j}{\sigma_{ij} \ \sum_{k \in V} w_k}$$

(7)

5.2.3. Example calculation

To give an example of its calculation, consider the simple networks in Figure 5.1. Note that the lengths of paths through the network are calculated metrically. Also, that the weights in this example, and henceforth, are the number of properties on each segment. This is because the number of properties on each segment are likely to better correlate with the quantity of origins and destinations than segment length used in other studies (e.g. Chiaradia, 2007; Turner, 2007). For one, because some longer roads (e.g. sections of motorways) are unlikely to represent origins or destinations despite their length and, vice versa, shorter segments may contain high density housing such as block of flats.

Figure 5.1a shows the network with equal weights or numbers of properties (one property) on each vertex in the network. Figure 5.1b shows the values from the standard
betweenness metric (Equation 5) for each vertex for which $x_1, x_2, x_3, x_5, x_6$ and $x_7$ receive the same value of 14 due to the seven routes that begin on those vertices combined with the seven routes from the other vertices that end at those vertices. In comparison, vertex $x_4$ receives a value of 22 – in addition to the 14 trips described above, it receives an extra 8 as the shortest routes between $x_1$ and $x_2$, $x_6$ and $x_7$ pass through it. As such, the standard betweenness measure would predict $x_4$ will experience the most movement and there will be equal amounts of movement on all other vertices. Note that because there are equal numbers of properties on each vertex, the adapted betweenness measure (Equation 7) will give the same betweenness values and so predict the same distributions of movement.

In comparison consider the very similar network in Figure 5.1c where there are equal numbers of properties (1) on $x_1, x_3, x_4, x_5$ and $x_6$ but there are no properties on $x_3$ and $x_7$. Because the standard betweenness measure (Equation 5) assumes all vertices contribute equally to the metric calculation it ignores the fact that some vertices may not function as an origin or destination, and so will estimate the same betweenness values as for the network in Figure 5.1a. The estimates are shown in Figure 5.1d. In reality, however, movement around the network would be more likely to resemble that shown in Figure 5.1e as calculated by the adapted betweenness measure (Equation 7). Here, vertices $x_2$ and $x_7$ will receive values of 0 as they cannot function as origins (no one can live on them), and for journeys to locations with homes, cannot function as destinations (as they contain no properties). Nor do they feature on any other routes through the network. Vertices $x_1, x_3, x_5$ and $x_6$ all receive values of 10 as they appear as origins to and destinations from to all 5 vertices and $x_4$ will receive a value of 12 as it will additionally feature on the shortest routes from $x_1$ to $x_6$ and from $x_6$ to $x_1$. 

115
Figure 5.1: Two example networks (a and c) and the associated betweenness values when using the original betweenness measure (b and d) and the adjusted betweenness measure (b and e).

(a) example road network

(b) betweenness score (estimated movement) from the original (and adjusted) betweenness measure

(c) second example network

(d) betweenness score (estimated movement) from the original betweenness measure

(e) betweenness score (estimated movement) from the adjusted betweenness measure

NOTE: In b and d, the predicted movement levels range from low (pale grey) to high (black).
5.2.4. **Guardians and guardianship**

A second key limitation with using these types of betweenness measures as an estimate of ambient guardianship is that the potential role of pedestrian passers-by is likely much more complex than simply a function of their presence or absence. To explain, and as described in the literature review in Chapter 2, according to several perspectives a person’s ability and willingness to act as a capable guardian may depend on several factors.

As described in Chapter 2, in Jane Jacobs’s (1961) *eyes on the street* perspective, it is argued that a person’s ability to act as a guardian is assumed constant in that guardianship is (equally) provided by everyone (not engaged in criminal activity) on a street (or road). In contrast, Oscar Newman’s (1972; see also Coleman, 1985) *defensible space* perspective argues whether the ambient population can deter crime depends on who is present. That is, people may not feel responsibility and therefore not act as guardians in areas outside their residential or local area. Also, if an area is used by many non-residents (non-locals) then the residents (locals) will be unable to be watchful of all outsiders and may not be able to recognise (and intervene in) suspicious circumstances. According to other theories (e.g. Felson, 1995; Reynald, (2011), a person’s likely capability to act as a guardian will depend on their attachment or proximity to an area. For example, a person will feel greater responsibility, and so will likely be more willing to act as a guardian, for properties or areas they intimately know (such as their own home or nearby) than for areas which they do not (such as, on average, those more distant from their house). Current methods of estimating betweenness however do not capture these nuances.

5.2.5. **Local and non-local betweenness**

To this end, a new set of betweenness-based measures are proposed to distinguish sections of people’s journeys that might be thought of as being more a part of their *local area*, those sections that are less so, those that are even less so, and so on if desired. In
Frith et al. (2017), this set of measures was introduced as a binary classification of local or non-local betweenness. In this case, the first sections of pedestrian trips from a person’s home location were classified as local movements, with the idea being that they would be most likely to act as capable guardians on these sections of road. All other parts of the pedestrian journeys were then assumed to be perceived as nonlocal and so will also act accordingly. Here, and for use in the analyses in Chapters 10 and 11, this set of measures is extended to allow for more nuanced testing of the above perspectives of guardianship (see also Discussion), to three bands for local (L), neither local or non-local (M), and non-local (N) movement. These are calculated using two thresholds, $l_1$ and $l_2$.

For the estimate of local guardianship ($L$), movement is only included and counted if it is from a segment with a property and the distance between the origin (which contains the property) and the focal vertex is less than $l_1$. Using Equation 7, the formula is therefore:

$$L = \sum_{i,j \in V, i \sim j, d_{ij} < l_1} \frac{\sigma_{ij}(e) w_i w_j}{\sigma_{ij} \sum_{k \in V} w_k}$$

For movement which is counted as neither local or non-local, M, movement is only included if the distances are greater than $l_1$ and less than $l_2$ and so the formula for $M$ is:

$$M = \sum_{i,j \in V, i \sim j, d_{ij} \geq l_1, d_{ij} < l_2} \frac{\sigma_{ij}(e) w_i w_j}{\sigma_{ij} \sum_{k \in V} w_k}$$

Lastly, for movement that is counted as non-local, N, movement is similarly only included but if the distances are greater than $l_2$:

$$N = \sum_{i,j \in V, i \sim j, d_{ij} \geq l_2} \frac{\sigma_{ij}(e) w_i w_j}{\sigma_{ij} \sum_{k \in V} w_k}$$

### 5.2.6. Example calculation

To give an example of its calculation, recall the hypothetical network in Figure 5.2a and the estimated betweenness values in Figure 5.2b (see also Figure 5.1a and 5.1b). Also, and
for simplicity imagine that only trips from \( x_1 \) are being calculated, that the travel time between each vertex and every other immediately connected vertex is five minutes, and that the two \textit{local} thresholds, \( l_1 \) and \( l_2 \), are five and ten minutes (travel time) respectively.

Using Equations 8 to 10 respectively, movement which is considered local is only on vertices closer than 5 minutes travel time away from vertices. As such, because it takes 5 minutes to reach any other vertex, either as the destination or on route to the destination vertex, the only vertex with local movement (from \( x_1 \)) is \( x_1 \). As illustrated in Figure 5.2b, it could be said (though only journeys from this vertex have been calculated) that there is the greatest potential for local movement (guardianship) at this vertex (\( x_1 \)).

Next, for movement to be considered neither local or non-local, only the sections of journeys which are between five and ten minutes away from vertices. In this case, \( x_2, x_3, x_4 \) and \( x_5 \) lie at these distances (from \( x_1 \)) on journeys through the network. However, because \( x_4 \) lies on more shortest paths through the network it receives a greater value than \( x_2, x_3 \) and \( x_5 \). Based on this, and as shown in Figure 5.7d, it could be said there is greatest potential for neither local or nonlocal movement (guardianship) on \( x_4 \).

Lastly, and again just calculated from \( x_1 \), nonlocal movement is on vertices which are over 10 minutes away. As such, only \( x_6 \) and \( x_7 \) are calculated to have any nonlocal movement as they are the only vertices at this distance (or more) from \( x_1 \) - all other vertices need at most five minutes travel time to reach. They are therefore the only vertices to experience this type of movement (guardianship) as illustrated in Figure 5.2e.
Figure 5.2: An example network (a) and the associated (overall) betweenness values (b), local betweenness (c), neither local or non-local (d) and non-local (e) values (calculated from x₁)

NOTE: In b-e, the predicted movement or guardianship levels range from low (pale grey) to high (black).
5.3. Modelling offender awareness spaces

5.3.1. Offender awareness spaces

While graph theory measures have been used in criminological research to estimate guardianship (or, as argued in section 5.3, to estimate the number of guardians present), graph theory can also be derived to estimate offender awareness spaces. That is, as described in Chapter 2, the areas where an offender likely frequents (such as their home, workplace, or local shops) or passes through (such as on the journeys between those locations) and so is more likely familiar with the crime opportunities. Knowledge of, or the ability to estimate, the awareness or familiarity of spaces is important as it is believed that offenders are more likely to, or prefer to, offend in familiar areas (see Chapter 2).

Despite its relevance to crime location research, previous studies have generally estimated familiarity using simplistic approaches. In most studies, this is done by measuring the distance to the offender’s home neighbourhood (or any other known activity nodes). This is based on the assumption that locations closer to the offender’s home (or other activity nodes) are more likely to be known and more familiar. This assumption not unreasonable, and is supported by findings from qualitative research with burglars, for example (e.g. Rengert and Wasilchick, 1985). However, an individual’s familiarity with locations will be more nuanced than this and will likely be more a function of how often they travel to or through those locations. As such, while distance will influence their awareness spaces, so too will the configuration of the road network as this affects the likelihood they will travel along a road to reach that or other locations (Beavon, Brantingham and Brantingham, 1994). For example, some roads (such as major roads) may feature on many journeys, while others (such as cul de sacs and dead ends) will less so and hence will differ in their level of familiarity to each offender. Furthermore, because the distance to reach a location (from the offender’s home) is also a measure of the effort needed (another key factor in
offence location choices), using distance as a measure of both (effort and familiarity) may conflate its observed effect and complicate any conclusions drawn from its impact.

Some studies also use other variables such as the distance to the city centre (e.g. Bernasco and Nieuwbeerta, 2005) and the presence of transport links to other areas (e.g. Clare, Fernandez and Morgan, 2009) as additional proxies for familiarity as these will also likely shape offender awareness spaces. For the former, this will be because of the concentration of facilities or properties (e.g. shops) in the city centre. As such, it is likely visited more often than other areas. The areas between it and the offender’s home will therefore be expected to be more familiar. For the latter, this will be because transport links enable the offender to reach specific other areas (where other transport stations are) with less effort. As such, they would be expected to visit those areas, more than other equivalent areas, and so will again be more familiar. If the transport system is above ground, they may also allow offenders to become aware of opportunities on the route. These effects are, however, likely more complicated than can be captured by these variables, and more intrinsically related to the shape of the road network and the distributions of properties. For example, if an offender lives closer to, or there is a greater concentration of facilities elsewhere than the city centre (e.g. at an out-of-town centre) then the offender will likely be more familiar with opportunities there and no-more familiar with those in city centre. Also, even if their neighbourhood has transport links to other areas, those areas will likely only be more visited (than before or compared to other areas) if the transport system meaningfully reduces the effort to reach them (e.g. compared to walking) and there are places there worth visiting there.

5.3.2. Idiosyncratic betweenness

Based on this, to incorporate the offender’s awareness space in analyses of offence location choices more accurately, a novel idiosyncratic version of the betweenness metric,
idiosyncratic betweenness, is proposed. To explain how it is derived, recall that (overall) betweenness is classically calculated using the following formula:

\[
B_e = \sum_{i,j \in V, i \neq j} \frac{\sigma_{ij}(e)}{\sigma_{ij}}
\]  

(11)

Which involves:

1. Taking a vertex in the network (\(i\))
2. Computing the shortest routes from this vertex to all (applicable) vertices in the network (\(j\))
3. Calculating the overlap of these routes on each vertex (\(e\)) (after accounting for multiple shortest routes to the same vertex)
4. Repeating step 1 to 3 for all other vertices in the network

This metric may be thought of as in the following way. For a person living on vertex \(i\), if they approximately visit all other vertices (\(j\)) in the network equally, the overlap in these routes (\(e\)) will be an estimate of how likely (or frequently) they would be expected to pass through or visit each vertex during their everyday urban activity (step 1-3). This is then repeated for journeys for all other vertices in the network (from all other vertices, \(j\)) (step 4) and the betweenness value estimates where it is expected that people, on average, converge with other people.

To estimate awareness spaces, this process does not need to be repeated from every vertex (step 4). Instead, betweenness (herein idiosyncratic betweenness) could be calculated from the one road segment or vertex where the offender lives using steps 1 to 3 above (or from the multiple segments or vertices where they have activity nodes if more are known). The calculated betweenness value for each vertex could therefore be interpreted as the
likelihood or frequency the offender will visit (and hence become aware of) a road segment. More formally, this is calculated for offender \( x \) whose activity nodes are \( y \) as:

\[
I^x_{\tilde{e}} = \sum_{j \in V, y \neq y \sim j} \frac{\sigma_{yj}(e)}{\sigma_{yj}}
\]  \hspace{1cm} (12)

Much like the standard betweenness measure, this idiosyncratic betweenness measure can also be modified to incorporate vertex-weighting (see section 5.2). That is, to account for the fact that some vertices are expected to be visited more or less often than others. Here, and based on Equations 7 and 12, this measure would be calculated as follows:

\[
I^x_{\tilde{e}} = \sum_{j \in V, y \neq y \sim j} \frac{\sigma_{yj}(e) w_j}{\sigma_{ij} \sum_{k \in V} w_k}
\]  \hspace{1cm} (13)

Here, because this metric is calculated for each offender (\( x \)) from their home (though it can be extended if multiple activity nodes are known), the origin weight, \( w_i \), is not needed as it will be 1 for journeys beginning from the vertex where they live and 0 for all others.

5.3.3. Example calculation

To give an example of its calculation, recall the hypothetical network from Figure 5.1a (see also Figure 5.3a) and the two scenarios in Figures 5.3b, 5.3c, 5.3d and 5.3e for which an offender lives on vertices \( x_1 \) and \( x_4 \), respectively. In this illustration, imagine that journeys are calculated to all other vertices, but not to the vertex on which the offender lives.

As illustrated in Figure 5.3b for an offender who lives on vertex \( x_1 \), their (shortest) routes to all other vertices in the network all involve traversing \( x_1 \) where they live while three routes (to \( x_4, x_6 \) and \( x_7 \)) involve traversing \( x_4 \). All other vertices only feature on one route. As such, as shown in Figure 5.3c, \( x_1 \) would receive the largest value (6) while \( x_4 \)
would receive the next highest value (3). All other vertices would equally receive a low value (1). This measure, from Equation 9, would therefore predict, as would be expected based on theory (see above), that the offender would be more likely to be familiar with opportunities on $x_1$ on which they live, followed by $x_4$ and then the remaining vertices. Also, and as would be expected, this is despite the fact that vertices $x_2$, $x_3$ and $x_5$ are the same distance from $x_1$ as $x_4$.

In contrast, as illustrated in Figure 5.3d, in the case where an offender lives on vertex $x_4$, their (shortest) routes to all vertices only involve traversing the offender’s home vertex ($x_4$) and the destination (vertex) of the journey. Therefore, and as shown in Figure 5.3e, $x_4$ would receive the largest value (i.e. 6, 1 for each route traversing some part of it) while all other vertices would equally receive a low value (1). Therefore, for this example, it is predicted that the offender would be very familiar with opportunities near their home (on $x_4$) but equally (poorly) familiar with opportunities elsewhere. Again, and although very different results are obtained compared to the first example, this is expected as if the offender does equally travel around this network, no other road segments (except for that where they live) need to be traversed (routinely) to reach other segments.

5.4. Discussion

In this chapter, three novel sets of betweenness-based graph theory measures were proposed for use in criminological research; and demonstrated using example networks. The first, vertex-weighted betweenness, builds upon the standard betweenness metric by accounting for the non-uniform distribution of origins and destinations through the network. It also builds on previous weighted betweenness measures for modelling movement by factoring in the fact that more possible destinations should not necessarily mean more journeys; but rather a greater choice of journeys.
Figure 5.3: An example network (a), the shortest routes to all other vertices from $x_1$ (b) and $x_4$ (d) and the corresponding idiosyncratic betweenness values (c and e)

(a) example road network

(b) the shortest paths to all other vertices for an offender living on $x_1$

(c) the calculated idiosyncratic betweenness value (familiarity) for an offender living on $x_1$

(d) the shortest paths to all other vertices for an offender living on $x_4$

(e) the calculated idiosyncratic betweenness value (familiarity) for an offender living on $x_4$

NOTE: Dashed lines represent multiple equal short routes; in c and e, the predicted movement or guardianship levels range from low (pale grey) to high (black).
The second set of measures incorporates several theories which argue that guardianship, and the ability of passers-by to deter crime, is likely related to who is on the street rather than just how many people there are. To this end, the local to nonlocal betweenness measures are computed by splitting the routes calculated as part of the betweenness measure into the sections which are closer or further away. These measures therefore attempt to measure the amount of the different movement types – in this analysis, of local, neither local or nonlocal, and nonlocal – rather than an estimate of the overall amount of movement.

The third measure attempts to measure an offender’s awareness space and to disentangle this from the effort associated with travelling to a particular location. These two factors are commonly conflated and estimated using the distance from the offender’s home. This new measure, idiosyncratic betweenness, accomplishes this by calculating the shortest routes (as calculated as part of any betweenness measure) to other vertices from only the vertex where the offender lives. Therefore, if the offender is assumed to travel around the network, the overlap in these routes represent where the offender is more likely to familiar and aware of the opportunities for crime. Because real-world networks are not uniform in shape the idiosyncratic betweenness will vary from distance (farness) and so the effects of effort (distance) and familiarity (idiosyncratic betweenness) can be, unlike before, disentangled.

While the proposed measures are based on the principles of graph theory and how the standard betweenness measure is calculated, used and interpreted in previous research (e.g. Hillier and Iida, 2005), the proposed measures are not, at present, quantitatively validated. This is in the sense that there are no experimental, or otherwise, research (Frith, Johnson and Fry, 2017) showing the metrics are truly estimating what they are supposed to be estimating. For example, that the awareness space estimates from the idiosyncratic
betweenness metric corresponds with offenders’ actual levels of familiarity (of crime opportunities) around a road network. Such studies are obviously needed to fully validate the proposed measures – but are beyond the scope of this chapter and thesis.

That said, as the first step in testing these proposed measures, they could be included in analyses, for example, of offence location choices, and their effects (on crime locations) could be compared against that predicted based on theory. For example, given the evidence that offenders tend to prefer offending in familiar areas (e.g. Rengert and Wasilchick, 1985), if greater levels of idiosyncratic betweenness are not associated with a greater risk of being selected for an offence, this suggests there is an issue with the idiosyncratic measure. To this end, several hypotheses regarding the metrics proposed in this chapter are developed and tested later in this thesis (in Chapters 10 and 11) after the specific types of analyses that will be used (spatial discrete choice analyses) are introduced and examined in relation to crime location choice research (in the next four chapters).
Discrete Choice Methods
Introduction to discrete choice methods

In this chapter, the four statistical approaches to analysing offence location choices and offender preferences are introduced. Key limitations of three of the approaches which have typically been used for these analyses are discussed. The fourth approach, the discrete choice approach, which overcomes these issues, but possesses its own shortcomings, is then explored. The three main discrete choice models (the conditional logit, mixed logit, and latent class) are then examined along with the relevant terminology and technical material.

6.1. Introduction

Until relatively recently, three main statistical approaches were used to analyse offence location choices and offender preferences. In the order of their appearance in the crime literature they can be described as the target-based, offender-based and mobility-based approaches.
6.2. Target-based approach

The target-based approach represents the earliest method (e.g. Balbi and Guerry, 1829) and the most common used in the published literature. Based on the now-established finding that crime spatially concentrates (e.g. Sherman, Gartin and Buerger, 1989), studies that adopt this approach attempt to describe variations in the volume or rate of offences at different locations as a function of the characteristics of those, or other relevant, locations. Such characteristics can refer to a range of ecological and environmental factors. The locations can also be defined using various units of analysis at different geographic scales. For example, Sampson (1985) examined the association between crime rates measured at the meso US neighbourhood level and the density of multi-unit housing, residential transiency, familial disorganisation, racial composition, unemployment and poverty. A second exemplar is that by Johnson and Bowers (2010) who used a multi-level model and micro-level street segments in Merseyside (UK) to examine the association between the risk of burglary and various socio-demographic influences (such as those above) and measures of street segment connectivity and area-level permeability. The key limitation with analyses using this approach is that they ignore (for exceptions see Bernasco and Luykx, 2003; Bernasco and Block, 2011) all information about offenders such as where they live, or they assume that distance is unimportant – which is at conflict with much of the research on offending (see below).

6.3. Offender-based approach

The second approach is the offender-based approach. Although this approach did not gain popularity until the 1970s, the first study of this kind was conducted by White in 1932. These studies attempt to examine the spatial distribution of offences relative to the offenders’ home (or other) locations. That is, using the offenders’ hypothesised journey-to-crime [JTC]. Although other aspects such as the directionality of the JTC have been
explored (e.g., Rengert, 1989), this research almost exclusively focuses on the length of the JTC and the distance the offender travels (from their home) to their offence locations (Rengert, 2004). A common finding is that of the distance decay pattern whereby most crime trips are observed to be short and the intensity of offending decreases with distance from offender home (or other) locations. The distance decay pattern is also hypothesised to contain a buffer space close to the offender’s home locations within which offending intensity is expected to be lower (e.g., Rossmo, 2000; Canter and Youngs, 2008). The rationale being that offenders will be more easily recognised close to their home and hence avoid such locations. However, there is little supportive evidence of this. Well known examples of these analyses include Wiles and Costello (2000) and Snook (2004) who, amongst other findings, also showed that the length of crime trips vary by offence type and the age of the offender. In contrast to the target-based approach, the key limitation with the offender-based approach is that it ignores all information on targets such as their attractiveness and location or assume that both are unimportant – which is also counter to much crime location research (see above).

6.4. Mobility-based approach

The third approach, first used by Smith (1976), is the mobility-based approach. Using spatial interaction or gravity models, the aim of this approach to analysis is to explain the flow of crime trips between pairs of locations as a function of: 1) the friction between them (such as distance); and, 2) the characteristics of the origin and destination locations that may generate (push) or attract (pull) offending respectively. In alternative derivations, friction is defined in terms of the number of intervening opportunities located at shorter distances (in any direction) than each destination (Stouffer, 1940) or specifically located between the origin and each destination (Stouffer, 1960). Smith (1976) and Elffers et al. (2008) found that spatial interaction models that used distance (rather than intervening opportunities)
provided a better fit to their data. However, such models have been used in very few studies. Two other examples are Reynald *et al.* (2008) and Peeters and Elffers (2010) who used them and found social and physical barriers (respectively) generate impedance in addition to distance. While mobility-based studies can relatively simply include both sets of information used in the target-based and offender-based approaches (i.e. variables concerning the offender’s relationship with the target and the target itself), they are themselves limited, for example, by requiring separate analyses for understanding the effects of offender-level variables (e.g. age and/or gender). Mobility-based analyses can therefore be impractical and in particular when considering that much qualitative (e.g. Bennett, Wright and Wright, 1984; Rengert and Wasilchick, 1985; Wright and Decker, 1996) and quantitative (e.g. Townsley and Sidebottom, 2010; Bouhana, Johnson and Porter, 2016) research suggest offenders substantially vary in their decision-making.

### 6.5. Discrete choice approach

To overcome this methodological gap, Bernasco and Nieuwbeerta in their now-seminal 2005 article, *How Do Residential Burglars Select Target Areas? A New Approach to the Analysis of Criminal Location Choice*, introduced the discrete choice approach to crime location choice research. The key advantage of this approach is that it can be used to analyse target choices and simultaneously include characteristics of the offenders (e.g. age), targets (e.g. attractiveness) and their relationship (e.g. proximity). Owing to this advantage, this approach has become a relatively popular approach and has now been used in over 20 further published studies (for a systematic literature review, see Chapter 9 of this thesis).

Originally developed within economics by econometrician Daniel McFadden (1974), the discrete choice approach is relatively straightforward to translate and apply to the analysis of offence location choices. That is, and generally following the notation elsewhere (e.g. Train, 2009), it considers there is a sample of decision-makers, \( n = 1, \ldots, N \). In the case
of the decision of where to offend, it is assumed the decision-makers are (suitably inclined) offenders. In deciding where to offend, each offender (decision-maker) faces a choice from \( j = 1, \ldots, J \) alternative locations (also called the choice set).

According to the theory underlying the approach, when choosing their offence location, it is assumed that each offender will choose the alternative that they expect to derive the most utility, \( U \). This is such that alternative \( i \) will be chosen if it is expected to yield the most utility: \( U_{ni} > U_{nj} \) \( \forall \ j \neq i \). By observing attributes of the alternatives, \( a_{nj} \), and the offender decision-maker, \( d_n \), a function can be specified regarding the decision maker’s utility: \( V_{nj} = V(a_{nj}, d_n) \) \( \forall \ j \). However as the utility that would be obtained from each alternative, \( U_{nj} \), is unknown (except to the offender themselves) and will probably not be fully described in any model, it is decomposed into \( V_{nj} \) which represents observed factors that influence this utility and \( \varepsilon_{nj} \) which captures unobserved factors: \( U_{nj} = V_{nj} + \varepsilon_{nj} \) \( \forall \ j \). This utility function can also be expressed as \( V_{nj} = \beta_n x_{nj} \) where \( x_{nj} \) is a vector of variables relating to each alternative and \( \beta_n \) is an associated vector of coefficients (preferences) to be estimated. Using this formulation, different empirical models can be generated based on different distributional specifications for \( \beta_n \) and \( \varepsilon_{nj} \).

### 6.5.1. Conditional logit

The simplest discrete choice model, and the one specifically introduced by Bernasco and Nieuwbeerta (2005) and used in most crime location choice studies (see also Chapter 9), is the conditional logit [CL] (McFadden, 1974). The CL assumes that \( \beta_n = \beta \) \( \forall \ n \) such that the preference for each attribute is identical across offenders, or if interacted with offender attributes such as age (e.g. as in Bernasco and Nieuwbeerta, 2005), only systematically differs based on that attribute. The CL also assumes that \( \varepsilon_{nj} \) are
independently and identically distributed with a Gumbel distribution. Together this gives the probability of offender \( n \) choosing neighbourhood \( i \) \( (P_{ni}) \) as:

\[
P_{ni} = \frac{\exp(\beta x_{ni})}{\sum_j \exp(\beta x_{nj})}
\]

(14)

Although the CL is popular, for example, because it can easily be computed due to its closed form, it is not without limitations. The first, and currently of most interest in crime location research regards \( \beta_n = \beta \forall n \) in that it is assumed that all offenders or sub-groups of offenders shared the same preferences for offence locations. Various research (see above) however suggests that offenders substantially vary in their decision-making. Although in some studies decision-maker heterogeneity has little overall effect on the results (e.g. Dahlberg and Eklöf, 2003; Persson, 2002), in others (e.g. Bhat, 2000; Revelt and Train, 1998) it can have a large impact. One particular problem with this is that estimates from the CL can result in coefficients that are smaller in magnitude than the true values (Revelt and Train, 1998). This occurs due to the normalisation of the extreme value term in the CL during the estimation process as this term incorporates all of the variance - including that between decision-makers (see also Chapter 9).

Secondly, the CL model implies strict substitution patterns including the independence of irrelevant alternatives [IIA] and an assumption that unobserved factors are independent over any repeated choices. Regarding the former, IIA implies that the relative odds of choosing any alternative over another does not depend on any other alternatives. While this seems reasonable, it ignores the fact that some alternatives are more (or less) similar to each other and so can be more (or less) substitutable than other alternatives. Lastly, regarding the latter, while state-dependence - where an offender’s earlier choices influence their future choices (in this analysis, the attribute related to previous offence location choices, see earlier)- can be accommodated, unobserved factors are assumed to be
unrelated. This can be an issue as it would be expected there are unobserved factors that influence offenders’ decisions and that these factors persist across choice occasions.

6.5.2. Mixed logit

At least in part due to these issues, researchers in other disciplines and more recently in criminology (Townsley et al., 2015b; Frith, Johnson and Fry, 2017), often specify more sophisticated models including the mixed logit [ML] model (McFadden and Train, 2000). Rather than fixed coefficients in CL, the ML accommodates unobserved preference heterogeneity by assuming preferences, $\beta_n$, follow a continuous probability density function, $f(\beta_n | \theta)$, where $\theta$ describes the distribution. This is such that $\beta$ can vary for each offender. ML can also be generalised to explicitly accommodate repeated choices (of offence locations) by the same offender where $\beta$ can still vary over offenders but is consistent across the (repeated) choice occasions for the same offender. This is such that conditional on $\beta_n$, the probability of offender $n$ choosing alternatives $i$ in choice occasions $i_1, \ldots, i_t$ is:

$$P_{ni} = \int \prod_t \left[ \frac{\exp(\beta x_{ni_t})}{\sum_j \exp(\beta x_{nj_t})} \right] f(\beta_n | \theta) \, d\beta_n$$  \hspace{1cm} (15)

Lastly, ML also does not exhibit IIA as the denominators in a ratio of two mixed logit probabilities are inside the integral and therefore do not cancel. The ratio therefore depends on all data, including all other alternatives. The ML however requires the specification of the preference distributions, albeit that normal or log-normal (if the sign of the preference is to be restricted) distributions are typically assumed (Hensher and Greene, 2003). A second weakness with the ML is that, unlike the CL, it has no closed form and so cannot be realistically solved analytically (Train, 2009). As such, the ML must
be estimated through simulation - with either maximum simulated likelihood [MSL] or hierarchical Bayes [HB] - which is computationally expensive.

6.5.3. Latent class logit

Alternatively, and although it can be considered a special case of the ML and is so-far unused in crime location choice research, unobserved heterogeneity can also be handled using the latent class logit [LC] model (Lazarsfeld and Henry, 1968; McLachlan and Peel, 2000). LC assumes that $\beta_n$ take a discrete distribution and that there are $c = 1, \ldots, C$ distinct sets (or classes) of $\beta_n$ and each offender belongs to each class with some probability. In other words, it assumes there are latent types of offenders where each type has a certain preference for each variable. This preference can differ across types but is identical within each type. Similar to the mixed logit, including the assumption that factors can influence offenders over $t$ choice occasions, the probability is given by:

$$P_{ni} = \sum_c \pi_{nc} \prod_t \left[ \frac{\exp(\beta_c x_{ni,t})}{\sum_j \exp(\beta_c x_{nj,t})} \right]$$ (16)

Where $\beta_c$ is the parameter for class $c$ and $\pi_{nc}$ is the probability that offender $n$ belongs to class $c$. When observable characteristics of the offenders are present and assumed to effect which type or class they belong to, $\pi_{nc}$ is given by:

$$\pi_{nc} = \frac{\exp(\delta_c z_n)}{\sum_j \exp(\delta_c z_n)}$$ (17)

Where $z_n$ are observable characteristics (such as gender) that are consistent within offenders and can affect the class membership of offender $n$ and $\delta_c$ is the associated parameter. In effect, the impact of observable characteristics on preferences can be investigated directly with the model without post-hoc analyses - as would be needed using the CL or ML. One limitation with the LC is that it requires specification of the number
of classes – though this is typically determined by repeatedly estimating models with varying numbers of classes and comparing information criteria values. Also, and in a similar way to the ML, another shortcoming of the LC is that as the number of parameters increase, computing it using MLE becomes computationally expensive and can fail to achieve convergence (Bhat, 1997; Train, 2008). As such, in most cases it is estimated either through maximum likelihood estimation (MLE) or expectation-maximisation (EM). That said, because MLE can take substantially longer to compute and can also fail to achieve convergence (Bhat, 1997; Train, 2008), they are often computed using the latter, EM (Dempster, Laird and Rubin, 1977), though with the caveat EM can converge to a locally maximum log-likelihood - rather than a global one. For this reason, it is recommended that LC models are estimated several times with different starting values and the solution with the highest likelihood selected.
Chapter 7: Comparison of the approaches to analysing offender location preferences

Chapter 6 introduced the four approaches to analysing offence locations and their limitations. This chapter aims to build on this and empirically assess and compare the effects of their shortcomings to demonstrate, or not, the efficacy of the more recently introduced and theoretically superior choice-based approach. This evaluation will use simulated burglary datasets so that the variables and their effects on offending are a priori known as they were used to generate the data. Each approach is then assessed by its ability to detect these effects.

7.1. Introduction

As introduced in Chapter 5, there are four approaches to analysing offence locations and offender spatial preferences: the target-based, offender-based, mobility-based and choice-based approaches. Although all four approaches are appropriate for testing hypotheses regarding patterns of offence locations, they can give contradictory results. This is because of the limitations of each type of analysis and how the statistical models can be specified. This is also because each approach is used to ask different questions of the data and so
the results should be interpreted slightly differently. More specifically, target-based analyses ask what effect does $x$ have on the count of offences (at a location)? Offender-based analyses ask what is relationship between proximity (between the offenders’ homes and the offence locations) and offending intensity. Mobility-based analyses ask what affect does $x$ have on the number of crime trips between two locations? Lastly, choice-based analyses ask what affect does $x$ have on the odds of an offender (who lives in a particular area) offending at a particular location? Nonetheless, each of these approaches can broadly be interpreted in terms of addressing the question: does $x$ influence offence locations? As such, they should give relatively similar results; at least in terms of the significance and sign of the effects of the same variables.

Based on this, this chapter will investigate and compare the results from analyses using the four approaches. Given the differences in how the results of each analysis can be interpreted, the frequency of correct and incorrect results, in terms of type I, II and III errors, are compared. This is possible as the analysed datasets are generated through simulations of a simplified offence location choice process. The correct result will therefore be known as it was used to create the synthetic datasets.

### 7.2. Comparisons of the four approaches to analysis

The key limitation of target-based analyses is that they only incorporate variables concerning the characteristics of potential target locations (alternative-specific variables) (for exceptions see Bernasco and Luykx, 2003; Bernasco and Block, 2011). They, therefore, ignore all information regarding the offenders, including where they live. In doing so, they fail to account for the ease with which offenders might access a location. If proximity (an individual-alternative specific variable) is an important determinant of where offenders offend, its omission from the analysis will lead to biased estimates of the influence of the other factors. More specifically, based on the first law of geography which
states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p.236), if offenders prefer to travel shorter distances, their offence locations will incidentally resemble where they live, or places they frequently visit. As such, locational characteristics associated with where offenders tend to live (see also below) will bias the estimates of the influence of these characteristics (and other collinear characteristics) on offence location choices. In the extreme, if propinquity is the only factor that determines offence locations, any associations detected between crime location choices and other factors will be entirely spurious. In this case, instead of providing insight into where offenders choose to offend, such analyses will only provide insight into where offenders are most likely to live (which is a different choice). In comparison, choice-based (and mobility-based) analyses can incorporate alternative-specific and individual-alternative specific variables and so jointly account for the influence of both types of variables (offender and target). Therefore, considering that offenders tend to reside in particular kinds of areas with location characteristics often associated (in either direction) with offending (and therefore included as variables in those analyses) (Shaw and McKay, 1942; Bruinsma et al., 2013):

Hypothesis 1: In cases where proximity is the only criterion offenders consider when selecting offence locations, but it is not included in the analysis (in target-based analyses compared to choice-based analyses), locational characteristics (often associated with offending) are more likely to be falsely detected as significant predictors of offending.

This weakness may also be apparent when using more realistic models of offence location choices, including when locational characteristics and proximity both influence offence locations. To illustrate, consider the results in Table 7.1. These show the findings for two negative binomial regressions of the characteristics associated with the areas within which
samples of burglars live. The first analysis is for a dataset of burglars living in Buckinghamshire (UK)\textsuperscript{14}. The second – which is used to illustrate that the Buckinghamshire data are not in some way unusual - is for a similar analysis reported in Bruinsma et al.’s (2013). Taking just the variables employed in both analyses, the two studies suggest burglars tend to live in areas with greater poverty or lower socioeconomic status whereas the analysis in Buckinghamshire also found they tended to live in areas with less ethnic heterogeneity and higher residential mobility. In comparison, in various analyses of offence locations (Bernasco and Nieuwbeerta, 2005; Townsley et al., 2015a; Frith, Johnson and Fry, 2017), these same variables are found, or at-least expected, to be positively associated with offending. One key result of this is that, for example with ethnic heterogeneity (but for any other variables that deviate between where offenders tend to live and where they prefer to offend), when offenders prefer to travel short distances,

\begin{table}
\centering
\caption{The estimated odds-ratios of where burglars live in Buckinghamshire and The Hague (Bruinsma et al., 2013).}
\begin{tabular}{lcc}
\hline
Variables & Buckinghamshire\textsuperscript{1} & The Hague\textsuperscript{2} \\
\hline
Ethnic heterogeneity (10\%) & 0.85** & 0.90 \\
Households (100) & 1.14** & - \\
Percentage detached (10\%) & 0.93** & - \\
Poverty (£50,000) / Socioeconomic status\textsuperscript{3} & 1.10** & 2.05** \\
Residential mobility (10\%) & 1.08** & 1.34** \\
\hline
\end{tabular}

* indicates the OR is significant at $p < 0.01$ and ** indicates the OR is significant at $p < 0.01$.

1 The data used, on 675 burglars living in Buckinghamshire, is the same data is that used in the analyses in Chapters 10 and 11.

2 Other models of offender rates (and crime rates) based on other versions of social disorganisation theory were also tested however the one presented is the most similar to that run on the burglars living in Buckinghamshire.

3 Although the socioeconomic status variable used in Bruinsma et al. (2013) is measured differently to the poverty variable in the Buckinghamshire model, it measures a similar concept. That said, the relatively large difference in effect between the two analyses (from 1.10 to 2.05) may be due to the difference in the scales of the two variables.

\textsuperscript{14} Note that this dataset is subsequently the focus in Chapters 10 and 11 and so is described in more detail there.
(because of the first law of geography) they are more likely to offend in areas with less diversity. However, offenders also simultaneously prefer to offend, and so are more likely to (all else being equal), in areas with greater diversity that require travelling further distances. When one of these variables (and there will be other examples than those discussed) is omitted from the analysis (in target-based analyses), the conflicting influences of those preferences can be estimated together or otherwise misestimated. In this case, the model will be more likely to misestimate its influence - either regarding its actual effect going undetected or potentially estimating its impact in the incorrect direction\textsuperscript{15}. This leads to:

\textit{Hypothesis 2}: When proximity and locational characteristics are criteria for selecting offence locations, but proximity is omitted from the analysis (in target-based analyses compared to choice-based analyses), the true effects of locational characteristics are more likely to go undetected or misestimated as significant negative predictors of offending.

In contrast to target-based analyses, offender-based analyses examine individual-alternative-specific variables (usually just distance) and so ignore all information regarding targets and the alternatives an offender could have selected but did not. This can bias the estimate of the role of distance on offending in two ways.

The first way regards the number of alternatives (i.e. the number of areas in terms of the spatial units of analysis). In a typical analysis of the journey-to-crime, the researcher calculates how many crime trips occur within a given radius (or statistical \textit{bin}) of the

\textsuperscript{15} The reverse is also possible where preferences for certain factors complement each other. In the example provided this is likely to occur for poverty and residential mobility and could result in an exaggerated estimate of the effects of those variables. This however cannot be tested here and compared across the approaches because it relies on the estimated effects in each approach to be at-least equivalent which is not true (see earlier).
Chapter 7: Comparison of the approaches to analysing offender location preferences

offenders’ home locations. They then produce a histogram that summarises this distribution. In doing so, they examine the observed journey to crime distances, but ignore the rate at which different distances would be expected, assuming each potential target area has some likelihood of being selected. To explain why this is important, consider that if the bin size is $r$, the first bin contains an area equal to $\pi r^2$. The second bin will contain a larger area equal to $(\pi \times 2r^2) - (\pi \times r^2)$, and so on. Each subsequent bin covers a larger area and so will likely contain more alternatives. Offences are therefore more likely in more distant bins. If the constraining effect of proximity is measured in this way (by the number of offences in each bin) then it will be underestimated as subsequent bins will have more offences than they should, relative to the true impact of propinquity. In effect, the distribution will appear more right-tailed than it should. In the extreme, if propinquity has no effect on offending this should create the appearance of a negative relationship between proximity and offending where offending is more common further from the offenders’ homes. The appearance of greater amounts of offending in subsequent bands in offender-based analyses may also create the appearance of a buffer space in the distance decay pattern where the peak distance for offending is not in the immediate vicinity (the first band) of the offender’s home. In contrast to this analytic approach, choice-based analyses statistically model the choices made and not made and, so incorporate any variation in the number of alternatives in each band. It is therefore expected:

**Hypothesis 3:** When there are no offence location choice criteria, but the number of potential alternatives at each distance is not incorporated in the analysis (in offender-based analyses compared to choice-based analyses), proximity is more likely to be falsely detected as a significant negative predictor of offending.

---

16 For example, if bin size is 1km, the first bin will contain 3.14km$^2$; the second: 9.42km$^2$ and the third: 15.71km$^2$ and so on where after the first bin each subsequent bin will contain an additional 6.28km$^2$. 

144
**Hypothesis 4:** When there are no offence location choice criteria, but the number of potential alternatives at each distance is not incorporated in the analysis (in offender-based analyses compared to choice-based analyses), buffer spaces are more likely to be falsely detected.

These effects are also likely to be exacerbated because offender-based analyses implicitly assume targets are uniform in nature. They, therefore, do not account for any influence that their characteristics might have on the distances offenders travel to offend. On the one hand, if offenders tend to live in areas with characteristics similar enough to those where they prefer to offend, even if proximity is unimportant they will incidentally tend to travel shorter distances as they are already (by living there) in the areas most attractive for offending. However, a more realistic scenario as suggested by the results in Bruinsma *et al.* (2013) (see also earlier) is that the areas where offenders, such as burglars, generally live only bear some similarity to those where they prefer to offend. As such, even if proximity is unimportant, offenders will travel some distance to find the most attractive areas for offending. In either case, distance has no real effect on offending, but by ignoring the other factors associated with offence locations, offender-based analyses may find a false (likely negative) relationship between proximity and offending. In the latter scenario, the upward bias of offending intensity in later bins relative to the true (no) effect of proximity may also cause the appearance of a buffer space. As discussed earlier, unlike offender-based analyses, choice-based analyses can incorporate alternative-specific (locational characteristics) and individual-specific (distance) variables and can jointly account for the effects of both types of variables. It is therefore expected:

**Hypothesis 5:** When locational characteristics are the only criteria for selecting offence locations but are not included in the analysis (in offender-based analyses compared
to choice-based analyses), proximity is more likely to be falsely detected as a significant predictor of offending.

*Hypothesis 6:* When locational characteristics are the only criteria for selecting offence locations but are not included in the analysis (in offender-based analyses compared to choice-based analyses), buffer spaces are more likely to be falsely detected.

These effects should also be found (and detected) in more realistic scenarios when distance and locational characteristics are both criteria for determining offence locations. This leads to:

*Hypothesis 7:* When proximity and locational characteristics are both criteria for selecting offence locations, the effect of proximity is more likely to be incorrectly detected (misestimated) in offender-based analyses than choice-based analyses.

*Hypothesis 8:* When proximity and locational characteristics are both criteria for selecting offence locations, buffer spaces are more likely to be falsely detected in offender-based analyses than choice-based analyses.

In contrast to offender-based analytic approaches, mobility-based analyses can incorporate all of the variables likely to influence offence location choices and hence might be expected to address the above issues. However, they employ aggregated data concerning the flow of crime trips from and to each location. As such, to investigate if and how subgroups (e.g. young and old offenders) differ regarding preferences, separate models must be estimated for each subgroup. Especially when the number of subgroups increases, this reduces the sample size available in each model which will diminish their statistical power. In comparison, choice-based analyses (using the basic conditional logit) can incorporate systematic variation between subgroups in the same model through interaction terms between attributes (of the offenders) which identify the subgroups and
any of the covariates in the choice model that are expected to vary over the subgroups. This allows for the more efficient use of the full sample when estimating the model. Therefore, and using an arbitrary offender attribute:

**Hypothesis 9:** When proximity and locational characteristics are criteria for selecting offence locations, and offending subgroups are analysed, compared to choice-based analyses, for mobility-based analyses the true effect of the variables are more likely to go undetected (due to statistical power) as significant predictors of offending.

### 7.3. Methodology

#### 7.3.1. Simulation process

To test these hypotheses a series of datasets were simulated for which the offender-offence decision making rule was known. The advantage of this approach is that rather than being restricted to identifying the frequency of contradictory results (where approaches give opposing results for the same variable), it is possible to determine the rate of *incorrect* results. This is because the *correct* result (regarding the direction and significance of the effect of a variable) is known as it was used when generating the data. In general, the four analytic approaches can only be meaningfully compared using this metric (whether the estimated effect of a variable is correctly detected or not) as the coefficients from each analysis have different interpretations and so cannot be directly compared (see also earlier). An *incorrect* result is therefore either: a) when the effect of a variable is falsely detected as significant, b) when the effect of a variable is falsely detected as not significant, or c) when the direction of the effect of a variable is incorrectly detected (and is significant). In other words, when there is a type I, type II or type III error (e.g. Mosteller, 1948; Neyman and Pearson, 1933).
Chapter 7: Comparison of the approaches to analysing offender location preferences

To create the simulated datasets, a sample of 675 (the number of detected offenders in the Buckinghamshire dataset) artificial burglars are first generated. To test hypothesis 9, which concerns the influence of offender characteristics, the artificial burglars are randomly assigned to one of two types (e.g. $a$ or $b$)\(^\text{17}\). Offenders assigned to group a prefer targets that are closer to their home locations, whereas those assigned to group b prefer those located further away. For all other hypotheses the offenders are identical.

Taking the unit of analysis as lower super output areas [LSOA] in Buckinghamshire (see also later), each artificial burglar is then probabilistically assigned a home location. This is based on the spatial distribution of the real apprehended burglars within the study area. By generating home locations in this way, the simulated burglars (and their offence location choices) should more reasonably resemble those in the real world. This is important, not just for ecological validity, but because many of the hypotheses are predicated on real-life empirical findings such as that offenders tend to live and offend in certain types of locations (see above). As such, any effects from this may not be found if the simulated offenders and their home locations were unrealistic (e.g. if they were randomly assigned a home location).

A behavioural rule is then specified for the simulated burglars in terms of multiplicative offence location preferences (see also below). For all but hypothesis 9, the burglars share the same behavioural rule. In hypothesis 9, the simulated burglars share the same offending preferences except for distance which reverses in direction between, but is identical within, the two subgroups. The offence location decision process is then simulated where the likelihood of a location being selected is determined by each

\(^{17}\)The type itself is arbitrary as it is only used to distinguish the artificial burglars into two subgroups to test the impact of analysing subgroups on the ability to detect effects (hypothesis 9). In practice, these subgroups differ in terms of one preferring closer targets and the other preferring more distant targets. Reasons for such differences could be the experience of the offender in that those which are less experienced are more likely to act impulsively than those with greater experience.
locations’ characteristics and the burglars’ preferences for those characteristics. This is such that the probability of a burglar choosing an area is given by the product of the burglars’ preference for each characteristic multiplied by the area’s amount of it. For example, consider a study area that contains two potential offence locations, \( a \) and \( b \). These locations only differ in terms of their amount of a characteristic. Location \( a \) has two units (of that characteristic) whereas location \( b \) has zero units. If a simulated burglar prefers locations with greater amounts of that characteristic to the extent that for every unit they are 50 per cent more likely (or 1.5 times more likely) to select that area (than any other area), then location \( a \) is \( 2.25 \) (\( 1 \times 1.5 \times 1.5 \)) times more likely to be selected than location \( b \). Or in other words, when generating the offence location choices, the offender has a 69 per cent probability of selecting location \( a \) and a 31 per cent probability of selecting location \( b \). To introduce the influence of chance, decision making is probabilistic rather than deterministic (see below).

For each hypothesis (except hypotheses 3 and 4 where the offence data are generated with no offending preferences) the above process is repeated using five effect sizes (strengths of preferences) for each of the variables (see below) that determine offence locations. These were used to test the extent that the results vary based on the eccentricity of preferences. Following the example earlier, the preferences can be described using odds-ratios where for every unit the odds of an area being selected increases (or decreases for the other type of burglars in hypothesis 9) multiplicatively by 1.05, 1.10, 1.25, 1.5 and 2. For example, in hypothesis 1 the odds of an area being selected increases by 5%, 10%, 25%, 50% or 100% for every unit of proximity (or for every kilometre closer to the offender’s home). All other criteria not noted in the hypotheses (including all criteria in hypotheses 3 and 4) do not affect offence location choices. In other words, those factors which are not stated to influence their offence location choices are ignored when generating that specific data.
Examples of the generated offender and offence datasets (containing 500 burglars/burglaries with burglaries generated using preferences for the locational characteristics and proximity of 1.5) are provided in Figure 7.1 which show some variation between the generated datasets.

For each hypothesis, each dataset is also generated using varying numbers of burglaries, 100, 250, 500, 750 and 1,000, to investigate the effect of dataset size on the reliability of the estimated coefficients. Lastly, for every hypothesis, a total of 2,500 datasets were generated so that no outlying simulated datasets dominate the results, and a summary of the findings across datasets produced. For hypotheses 1, 2 and 5-9: the 2,500 datasets were created by generating 100 replications for each combination of the five effect sizes and five dataset sizes. Because there are no offending criteria for hypotheses 3 and 4, there is only one effect size (1 or no effect), and so the 2,500 datasets were created by generating 500 replications for each of five dataset sizes.

7.3.2. Data

The study area for these analyses comprises the county of Buckinghamshire (UK). This study was selected due to the availability of burglary data on a sample of (675) burglars who resided within the study area (between April 2004 and March 2014). The spatial units used in this analysis, in terms of the alternatives or the potential locations offenders live or select targets from were U.K. census Lower Super Output Areas [LSOAs]. These were selected on the basis that they represent relatively homogeneous spatial areas. They also likely reflect how residents define neighbourhoods (e.g. Sutherland, Brunton-Smith and Jackson, 2013). There are 319 LSOAs in the study area.

Distances (as used in the analyses of the crime trip lengths that offenders travelled or must travel to offend) were calculated as the length of the path along the road network using the Ordnance Survey highways road network dataset. Although some studies use
Figure 7.1: Examples of the simulated offender and offence datasets

Distributions of where the simulated offenders live

Distributions of the simulated crime trips

Distributions of the simulated offence locations
Chapter 7: Comparison of the approaches to analysing offender location preferences

*Euclidean* (straight line) distance, and others suggest little difference between this and road network distance (e.g. Townsley and Sidebottom, 2010), Euclidean distance is rejected as it neglects the role of the configuration of the road network in determining actual path lengths (Rossmo, Davies and Patrick, 2004). Inter-LSOA distances were calculated as the distance between each pair of LSOA centroids, and intra-LSOA distances were calculated as the mean distance between every street segment within each LSOA. Distances are then inverted (to form proximity) and collapsed into 2km bins18.

The ecological and environmental influences of burglary locations were incorporated using the 2011 UK census and 2006-2016 UK Valuation Office Agency [VOA] data. These variables (see Table 7.2) are not the primary focus of this analysis, but were selected due to their general use in related studies and particularly in Bernasco and Nieuwbeerta (2005). These include ethnic heterogeneity, the percentage of houses that are detached (which is equivalent to Bernasco and Nieuwbeerta’s *single-family dwelling* variable) and residential transiency (all calculated from the Census) and poverty (computed from the VOA). Further information on the derivation and summary information of these variables are provided in Table 7.2.

---

18 When hypotheses are tested by comparing the distance-decay patterns, only the first 10 bands are included. This is for two reasons. Firstly, the non-linear effects of proximity are hypothesised to (at-least partly) occur due to the increasing numbers of alternatives at greater distances from the offenders’ homes. However, because the study area is truncated by the boundaries of Buckinghamshire, at some distance (e.g. 20km) the number of alternatives will artificially stop increasing (and eventually decrease) which will affect the results. Secondly, to detect non-linear effects, such as buffer spaces, in choice-based analyses either dummy variables representing proximity need to be included or some transformation(s) of proximity. While the former allows the data to speak for itself and so is used, the latter requires those transformations to be specified and (likely) compared to others to determine the approximate shape. That said, including dummy variables for all of the (40) 2km bands would be infeasible statistically in all but the largest dataset sizes and so a cut-off, such as 10 bands as identified before, must be used. Those beyond this cut-off (20km+) are treated as belonging to a single band but are omitted in any subsequent analyses of the distance-decay shape (e.g. of the relationship between proximity and offending).
Chapter 7: Comparison of the approaches to analysing offender location preferences

7.3.3. Analytic methodology

The simulated offence datasets are then subject to various analyses. All of which are calculated using built-in commands in Stata (StataCorp, 2015). For the target-based analyses, the relationship between the number of burglaries in each LSOA and its locational characteristics (see Table 7.2) were estimated using count-based regression models. This type of statistical model was used as numbers of burglaries at the micro-level represent relatively rare events and tended to concentrate across LSOAs to approximately form a Poisson distribution. Although Poisson regressions could be estimated, negative binomial regression [NBR] models are favoured and so are estimated. This is because unlike Poisson regressions, NBRs add an over-dispersion parameter that can account for the existence of unexplained variability (between LSOAs) (Greene, 1994). If the dispersion parameters are close to zero, the NBR approximates the Poisson model, so there is no disadvantage to its application.

Table 7.2: Summary information of the independent variables used in the analyses

<table>
<thead>
<tr>
<th>Variable</th>
<th>Derivation</th>
<th>Mean</th>
<th>S.D.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethnic heterogeneity</td>
<td>The index of qualitative variation calculated using white, black, Asian (homogenous categories) and other (heterogeneous)</td>
<td>7.9</td>
<td>1.6</td>
<td>3.7</td>
<td>9.7</td>
</tr>
<tr>
<td>Households (100)</td>
<td>The number of households</td>
<td>6.5</td>
<td>1.3</td>
<td>4.3</td>
<td>11.9</td>
</tr>
<tr>
<td>Percentage detached</td>
<td>The percentage of households that are detached</td>
<td>3.6</td>
<td>2.4</td>
<td>0.2</td>
<td>9.7</td>
</tr>
<tr>
<td>Poverty (£50,000)</td>
<td>The (additive) inverse of the average median price of households sold</td>
<td>-4.9</td>
<td>2.2</td>
<td>-14.6</td>
<td>-1.7</td>
</tr>
<tr>
<td>Residential mobility</td>
<td>The index of qualitative variation calculated using residents in the same area one year ago (homogenous) and not (heterogeneous)</td>
<td>8.1</td>
<td>0.8</td>
<td>4.4</td>
<td>9.2</td>
</tr>
</tbody>
</table>

Note: All variables are scaled such that one-unit increases in the variable will be associated with non-negligible influences on offence locations. All variables are also derived so that according to theory and previous studies they should have a positive impact on offence locations.
Chapter 7: Comparison of the approaches to analysing offender location preferences

For the offender-based analyses, the first part of the conventional JTC methodology was followed where the distribution of crime trip lengths is split into ranges (bins) and represented by the frequency of crime trips within each bin (i.e. as a histogram). For this analysis, various bin sizes were tested (1km, 2km and 5km) but the analyses were sensitive to or obscured by the smaller and larger bin sizes respectively. As such, bin sizes of 2km were used. In typical analyses of the resulting histogram researchers often simply observe its shape (e.g. Canter and Youngs, 2008) or fit a mathematical function (e.g. Kent, Leitner and Curtis, 2006). Here a simpler (compared to the latter) but testable (unlike the former) approach is taken. This is because only two aspects of the distance-decay shape are important in these analyses: the general relationship between proximity and offending intensity and the presence (or absence) of buffer spaces. As such, Spearman’s correlation, between the number of crimes in each band and the distance interval of the bands, is used to identify any systematic effect of proximity (in hypotheses 3 and 5). For hypothesis 7, the shape of the distance-decay pattern is compared to that expected by the underlying preference for proximity (that was used to generate the data) using a chi-square goodness of fit test. Exact chi-square tests are used to identify buffer spaces (hypotheses 4, 6, and 8) by comparing the frequency of offences in the first band and the subsequent band with the greatest number of offences. If they significantly differ, and the subsequent band has a greater frequency of burglaries, a buffer space is present.

For the mobility-based analyses, burglary flows between LSOAs are modelled using count-based regression models. These are used rather than the log-linear specification employed in previous criminological studies (e.g. Elffers et al., 2008) for several reasons. For example, flows are frequently zero which cannot be log-linearised. The standard countermeasures to this such as truncating zero flows or adding a small positive number can lead to non-negligibly biased estimates (see also Burger, Van Oort and Linders, 2009).
Furthermore, for the same reasons as for the target-based analyses, NBR (gravity) models are estimated. Here the number of crime trips between each pair of origins and destinations is the dependent variable. The explanatory variables are the locational characteristics of each destination (see earlier), the distance between each pair of locations, and like other similar analyses, the number of cleared burglaries from each origin (outflow) and the number of cleared burglaries to each destination (inflow). The inflow variable also acts as a proxy for other unmeasured characteristics that make locations attractive to burglars. For the analysis of offending subgroups (see hypothesis 6), separate models are estimated for the two subgroups.

For the choice-based analyses, the popular and basic conditional logit model is used which requires three elements to be defined. These are the actor making each location choice (the individual burglars), the alternatives they are choosing between (LSOAs) and the criteria used to choose between the alternatives (the variables outlined earlier). The model assumes the *independence of irrelevant alternatives* [IIA] where the choice between two alternatives is independent of all others. IIA can be violated when the alternatives are similar. This assumption was tested by applying the Hausman-McFadden (1984) test on ten percent of the synthetic datasets, and no significant issues were identified. To compare the results regarding proximity and buffer spaces with the offender-based analyses (hypotheses 2-5) proximity is entered as a dummy variable and the predicted frequencies (after the statistical model accounts for all other influences) of offences within each proximity band are calculated. The distance decay pattern is then analysed using the same approach as that used for the offender-based approach results (see also earlier). This approach is taken so that the detection of any effect of proximity, particularly any nonlinear effect such as buffer spaces, is identified through a comparable approach to that used for the offender-based results. As such, any differences can be attributed to the
approach to analysis. For all other analyses, standard practice (e.g. Bernasco and Nieuwbeerta, 2005) is followed where proximity is entered as a continuous covariate.

To test hypotheses, McNemar’s tests (1947) and its clustered equivalent, Yang’s modified-Obuchowski [YMO] tests (Yang, Sun and Hardin, 2010; see also Obuchowski, 1998), are used. In this case, to determine if there are significant differences in the number of (type I, II and/or III) errors between the compared analyses. These tests are used here because the data are paired. To explain, the units of analysis are each of the variables/effects (related to the hypotheses) in every dataset and the response of interest is whether an error (type I, II and/or III) is found. Therefore, each unit of analysis has paired observations - whether an error is discovered in either of the two compared analyses - and so Pearson’s chi-square tests (1900) are not appropriate. For hypotheses 3-8 each analysis provides a single observation (the effect of proximity or presence/absence of buffer spaces) from each dataset and so the McNemar’s tests can be used. For the other hypotheses, each analysis provides multiple observations (for each locational characteristic and proximity) from each dataset and so the observations are clustered within those datasets. As such, McNemar’s tests are not appropriate, and the YMO test is used and preferred over other similar tests due to its general superior performance and because it is free of assumptions regarding the structure of the correlation in the data (Yang, Sun and Hardin, 2010). These tests were calculated using the ‘clust.bin.pair’ package (Gopstein, 2016) in the statistical software environment R (R Core Team, 2016).

7.4. Results

7.4.1. Target-based analyses

The first two hypotheses concern target-based (and choice-based) analyses. The first hypothesis predicted that when proximity is the only factor determining offence locations, errors in terms of locational characteristics falsely detected as significant predictors of
offending are more likely in target-based analyses than choice-based analyses. The associated YMO test which compares the proportion of errors in the two analyses of the same simulated datasets supports this prediction ($p < 0.01$). Specifically, only 5% of the locational characteristic variables (across datasets) were falsely detected as significant predictors in the choice-based analyses compared to 25% in the target-based analyses.

As shown in Figure 7.2, the proportion of errors is relatively consistent in the choice-based analyses across locational characteristics (between 4% and 6%) and when the datasets are generated with different numbers of burglaries and effect sizes (of proximity). In comparison, they substantially vary for the target-based analyses across the locational characteristics (from 6% to 73% for the ethnic heterogeneity variable) and increase with larger datasets (from 16% to 31%) and effect sizes (from 11% to 37%). This variation in the frequency of errors across locational characteristic variables in the target-based analyses generally follows from what was expected in that the locational characteristics

**Figure 7.2:** The percentage of locational characteristic variables falsely detected as significant predictors of offending by locational characteristic, dataset size and effect size (of proximity) in target-based and choice-based analyses (hypothesis 1)
most associated with where offenders live (ethnic heterogeneity and poverty; see also Table 7.1) are most likely to falsely appear as predictors of offending. The exception to this is the ‘number of households’ variable. This though can be explained by the fact that unlike the variables above and that expected by Tobler’s law (1970); it does not substantially spatially auto-correlate\(^{19}\). In other words, LSOAs with high or low numbers of households do not cluster. As such, when proximity is the only factor determining offence locations, those locations near an offender’s home where they are most likely to offend due to their proximity are not much more likely to also contain high or low numbers of households than those further away.

Regarding dataset and effect sizes, the results are also as expected. The ability to statistically detect effects, which in these analyses are falsely significant locational characteristic variables caused by the effect of proximity, depends on statistical power. This increases with larger datasets and when the apparent effect of a variable (which will appear to increase when the effect size of proximity is larger as offenders are more likely to offend in closer locations) is greater.

For the second hypothesis, proximity and locational characteristics both determine offence locations. However, it is expected that the true (positive) effects of locational characteristics on offending are more likely to be misestimated (either as not significant or as negative predictors) in target-based analyses than in choice-based analyses. The

---

\(^{19}\)One way to measure spatial autocorrelation is using the Moran’s I statistic (1950) which ranges from -1 when areas with high and low values are perfectly spatially dispersed (e.g. like a chess board), 0 when those areas are equally clustered and dispersed (i.e. randomly located) and 1 when they are perfectly clustered. Using this, the ‘number of households’ is slightly ($I = 0.03$), but significantly ($p < 0.05$) clustered whereas ethnic heterogeneity ($I = 0.69, p < 0.01$) and poverty ($I = 0.45, p < 0.01$) are both substantially and significantly clustered. For interest, the percent detached is also substantially and significantly auto-correlated ($I = 0.36, p < 0.01$) while residential mobility ($I = 0.04, p < 0.01$) is slightly but significantly auto-correlated.
significant YMO test of this hypothesis supports this prediction \( p < 0.01 \) where 27% of the effects of locational characteristic variables were incorrectly detected in choice-based analyses compared to 47% of the same variables in the target-based analyses.

As shown in Figure 7.3, the percentage of errors for both analytic approaches again vary across the locational characteristic variables, being between 21% and 45% for the choice-based analyses and between 27% and 99% for the target-based analyses. Although the large number of errors for ethnic heterogeneity (in target-based analyses) can be explained by its high spatial auto-correlation (see above), the relatively higher numbers of errors for residential mobility (in both types of analyses) cannot be explained by the same mechanism. This increase though may be explained by the distribution of the variable, in terms of its low dispersion (see also Table 2) across the LSOAs. In more detail, if the values of one variable (e.g. residential mobility) across LSOAs are all similar (close to the

Figure 7.3: The percentage of locational characteristic variables undetected as significant positive predictors of offending by locational characteristic, dataset size and effect size (of all variables) in target-based and choice-based analyses (hypothesis 2)
mean), the statistical models may be unable to recognise its impact, in this analysis its positive impact, on offending.

For both approaches to analysis, the percentage of errors decreases as datasets and effect sizes become larger. In a similar way to hypothesis 1, this is to be expected as the errors (likely to be) found in these analyses are where the variables are not detected as significant, and these are more likely in smaller datasets (as the statistics have less power) and when the underlying effect size is small and does not substantially differ from no effect.

7.4.2. Offender-based analyses

Hypotheses 3-8 all concern offender-based (and choice-based) analyses and predict the distributions of the distances offenders travel to offend are more likely to show incorrect results (errors) in offender-based analyses than choice-based analyses. In particular, they are more liable to find negative relationships between proximity and offending (hypotheses 3 and 5), misestimate the constant multiplicative effect of proximity (hypothesis 7) or falsely detect buffer spaces (hypotheses 4, 6 and 8).

Hypotheses 3 and 4 were tested using simulated offenders that had no offending preferences. As such, the distance-decay pattern should appear as a straight horizontal line as offences are equally likely in each distance band. Shown in the top panel of Figure 7.4 are the distributions of offences over the first ten distance bands for two example simulated datasets, and in the bottom panel, the average across all datasets according to both approaches to analysis. From a visual inspection, the distributions are somewhat similar. However, further inspection shows that it is clear that all three distributions appear to show an overall positive relationship between offending and proximity when using offender-based analyses, but not when using choice-based analyses. Similarly, buffer spaces also seem to be present, or at least more prominent, in the offender-based analyses than the choice-based analyses. Statistical analyses of these distributions support these
observations with none of these errors found in the choice-based distributions while buffer spaces were found in all three offender-based distributions and positive relationships (between offending and proximity) were detected in the first example dataset and the average dataset. A significant McNemar’s test of the results across all of the generated datasets support these findings. More specifically, a negative relationship was found in 4% of the datasets when subjected to choice-based analyses compared to in 37% of the same datasets when subjected to offender-based analyses ($p < 0.01$). Similarly, buffer spaces were found in 35% of the datasets when subjected to choice-based analyses compared to 88% of the datasets when using the offender-based analyses ($p < 0.01$).
As shown in Figure 7.5, both types of errors were also generally more frequent for larger datasets. For example, buffer spaces were found in every dataset generated with 500-1,000 burglaries when subjected to offender-based analyses.

Hypotheses 5 and 6 test whether the consideration of locational characteristics (but not proximity) leads to errors of inference regarding the role of proximity on target choice. Shown in the top panel of Figure 7.6 are the distributions of offences for two new example simulated datasets (using this offence location criterion), and in the bottom panel, the average across all simulated datasets, when subject to offender-based and choice-based analyses. From a visual analysis of the offender-based results, there (again) appears to be negative relationships between proximity and offending and buffer spaces. In comparison, the patterns for the choice-based analyses appear to (correctly) indicate no (systematic) relationship between proximity and offending, though there may be a slight positive correlation in the second example dataset. They also seem to suggest the absence of buffer spaces correctly. The statistical analyses of these distributions find the same results and as do the McNemar’s test ($p < 0.01$) comparing the results from all 2,500 generated datasets. In particular, significant (negative or positive) relationships between proximity and offending were far more likely to be found in the offender-based analyses.

**Figure 7.5:** The percentage of errors found, in terms of negative relationships between proximity and offending (hypothesis 3) and the presence of buffer spaces (hypothesis 4), in the distance decay patterns from offender-based and choice-based analyses
analyses (in 73% of the analyses) than in choice-based analyses (in 7%). Likewise, buffer spaces were also more likely to be found in offender-based analyses (in 95% of the analyses) than in choice-based analyses (in 42%).

The frequency of these errors in datasets generated with different numbers of burglaries and effect sizes are shown in Figure 7.7. Of note, errors are more common in offender-based analyses when using larger datasets including where buffer spaces were found in 97% of the datasets generated with 250 burglaries and 100% of the datasets generated with 500, 750 or 1,000 burglaries. In comparison, the frequency of errors in choice-based analyses remains relatively consistent.

For the final hypotheses, 7 and 8, regarding offender-based analyses, a more realistic model of offence location choices was simulated. Here, the artificial offenders selected
Chapter 7: Comparison of the approaches to analysing offender location preferences

Figure 7.7: The percentage of errors found, in terms of significant relationships between proximity and offending (hypothesis 5) and the presence of buffer spaces (hypothesis 6), in the distance decay patterns from offender-based and choice-based analyses

offence locations by considering the locational characteristics and proximity of each potential target location. Shown in the top panel of Figure 7.8 are the distributions of offences for two new example simulated datasets (using this offence location criterion), and in the bottom panel, the average across all the simulated datasets when subjected to offender-based and choice-based analyses. From a visual inspection of these distributions, it appears the distributions from offender-based analyses are substantially different from those expected by the effect of proximity, including where buffer spaces seem to be present in all datasets. In comparison, the distributions from choice-based analyses are very similar to those expected, and no buffer spaces appear to be present. The statistical analyses of these distributions generally support these qualitative findings though they fail to detect any buffer space in the second example dataset as the numbers of offences do not differ enough. The McNemar’s tests comparing the presence of both errors across 2,500 datasets are also both significant ($p < 0.01$) where the distributions of offences
Chapter 7: Comparison of the approaches to analysing offender location preferences

Figure 7.8: The percentage of offences across the first 10 distance bands in two example datasets (top) and the average across all datasets (bottom) using offender-based and choice-based analyses (hypotheses 7 and 8)

significantly differ from that expected in 72% of the target-based analyses compared to 2% of the choice-based analyses. Also, buffer spaces were found in 97% of the target-based analysis distributions compared to 13% of the choice-based analyses.

As shown in Figure 7.9, the proportion of distributions with errors also vary to some degree when produced from different sized datasets and datasets generated with different effect sizes. Of note, buffer spaces are more likely to be found in target-based analyses when the datasets contain greater numbers of offences. These also become less frequent (in both types of analyses) when larger effect sizes (for all of the variables) are used. Regarding the shape of the distance-decay pattern compared to that expected by the influence of proximity, these errors are present in almost all target-based analyses so no
trend can be substantiated. For choice-based analyses, these errors appear to be relatively consistent across dataset and effect sizes\(^20\).

### 7.4.3. Mobility-based analyses

The final hypothesis of this section, hypothesis 9, concerns mobility-based (and choice-based) analyses. This hypothesis predicted that when offending subgroups are analysed using mobility-based analyses, because separate models must be estimated for each subset, the effects of variables is more likely to go undetected (as significant predictors of offending). The YMO tests of this which compare the frequency of errors (any variables going undetected) in the choice-based analyses to the mobility-based analyses of the same simulated datasets supports this prediction \((p < 0.01)\). In particular, in the choice-based

\[^20\text{Note that in these analyses effect size refers to that for all variables (proximity and locational characteristics). When the effect size for proximity is larger than for the locational characteristics buffer spaces become less frequent.}\]
analyses only 20% of the variables (across all datasets) were falsely found to be non-significant predictors of offending compared to 51% (or 50% and 51% in each of the models of the two subgroups) of the variables when using mobility-based analyses.

As shown in Figure 7.10, in the choice-based analyses similar proportions (22% to 25%) of four of the six variables went undetected as significant predictors of offending. The exception to this are residential mobility (see earlier) and proximity which was detected in all but 1 of the analyses. Overall, relatively similar proportions (43%-60%) of all variables (excluding the distance variable) went undetected as significant predictors in the mobility-based analyses.

As also shown in Figure 7.10, but as would be expected, in both approaches to analysis, the proportion of variables that go undetected decreases as datasets and effect sizes (of all variables) increase. Regarding the dataset size, it is evident from Figure 7.10 that the choice-based analysis requires far fewer burglaries to estimate the effects of the variables.

**Figure 7.10:** The percentage of variables that go undetected as significant positive/negative predictors of offending by the variable, dataset size and effect size in mobility-based and choice-based analyses (hypothesis 9)
correctly. This is true even when considering that each mobility-based analysis has half the sample (as the sample must be divided into the subgroups and models run for each) as that used in the equivalent choice-based analysis. For example, using YMO tests, there are still a significantly greater proportion of variables which are not correctly detected as significant predictors of offending in the two mobility-based analyses using 1,000 burglaries than the choice-based analysis using 100 burglaries ($p < 0.01$).

7.5. **Discussion**

The aim of the analyses presented in this chapter were to compare the accuracy of the four analytic approaches commonly used in the literature to examine offender location preferences. To test their accuracy and susceptibility to biases, the four approaches were used to analyse synthetic datasets for which the data generating process was known. The expectation was that the choice-based approach would be the most accurate, and that the alternative approaches would be more likely to detect the influence of factors that were not involved in the generation of the synthetic data.

The analyses conducted to test hypotheses 1 and 2 showed that typical target-based analyses, which omit the role of proximity, can falsely detect certain variables as influencing offending patterns even when they do not. Similarly, the analyses for hypotheses 3-8 showed that the environmental backdrop can cause offender-based analyses to suggest incorrect relationships between proximity and offending even when they do not exist. This includes falsely detecting buffer-like effects. For example, although different bands are used, Figure 7.11 shows the buffer spaces observed in published studies using the offender-based approach (e.g. Canter and Youngs, 2008; Rossmo, 2000) are not unlike those which (falsely) emerge for analyses conducted using the same analytic approach when no such pattern actually exists. Lastly, the analyses also showed that when modelling limited systematic differences between subgroups (i.e. in the case of systematic
differences in one variable), choice-based analyses are more efficient than the equivalent mobility-based analyses.

In summary so-far, these analyses suggest that, all else being equal, that choice-based analyses are more likely to detect the effects of variables correctly than the other approaches. One weakness with these analyses is that they assume that equally sized and biased datasets would be available for all four types of analysis. In reality, the different types of analysis require different data and that data can differ based on attrition (see also Chapter 8). For example, because target-based analyses only require reported crime while the other approaches require detected crime. The effects of this are examined in Chapter 8.
Appendix A. Multi-level mixed-effects Bernoulli regression of the clearance rates of crime

To estimate the influence of possible factors on the likelihood of an offence being cleared, an approach similar to that used in Paré et al. (2007) is employed. More specifically, a multi-level mixed-effects Bernoulli regression is estimated where burglaries (level 1) are clustered within LSOAs (level 2). The dependent variable, whether a crime is cleared or not, is measured at the incident (or burglary) level. These burglaries are derived from police.uk open data and whether they are cleared or not cleared is calculated using the criteria in Table 7.3 below. Between 2013 and 2016 there were 10,220 burglaries in the study area. Of these 353 were omitted because of the outcome, in terms of whether an offender is arrested or not, is unclear.

For consistency with the analyses in the main part of this chapter and because police.uk data is relatively accurate, particularly for burglary, at this spatial unit (Tompson et al., 2015) and so is appropriate for defining the clusters, the clusters and the seven explanatory variables are measured at the LSOA level. Four of these variables are included in this analysis on the basis of their use in the main part of this chapter. That said, they are also hypothesised to affect burglary clearance rates in the following ways:

- Firstly, and also in line with Paré et al. (2007), poverty is likely to affect crime clearance rates in two contrasting ways. The first follows that the police are known to target and patrol poorer communities more thoroughly than those that are more affluent areas.

---

21 Note that the Bernoulli regression should generate approximately the same results as the equivalent logistic regression.

22 For comparison, equivalent analyses of whether a burglary is cleared or not but using all burglaries in the UK rather than in the study area is also presented. Note that due to the limited availability of police station location data, the police workload and distance to the nearest police station variables (which require this data) are omitted from these analyses.
The alternative argument is that the police are also known to discriminate against low-status victims (such as those that are poorer) and so may not be as thorough in their investigations of crimes against these victims. The evidence for either hypothesised effect, however, is mixed.

Table 7.3: List of police.uk crime outcome categories and their classification as ‘cleared’, ‘not cleared’ or ‘N/A’

<table>
<thead>
<tr>
<th>Police.uk outcome</th>
<th>Cleared Outcomes</th>
<th>N (Study Area)</th>
<th>N (UK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action to be taken by another organisation</td>
<td>Cleared</td>
<td>1</td>
<td>114</td>
</tr>
<tr>
<td>Awaiting court outcome</td>
<td>Cleared</td>
<td>23</td>
<td>6,107</td>
</tr>
<tr>
<td>Court case unable to proceed</td>
<td>Cleared</td>
<td>42</td>
<td>4,894</td>
</tr>
<tr>
<td>Court result unavailable</td>
<td>Cleared</td>
<td>183</td>
<td>33,524</td>
</tr>
<tr>
<td>Defendant found not guilty</td>
<td>Cleared</td>
<td>110</td>
<td>12,496</td>
</tr>
<tr>
<td>Defendant sent to Crown Court</td>
<td>Cleared</td>
<td>2</td>
<td>123</td>
</tr>
<tr>
<td>Formal action is not in the public interest</td>
<td>Cleared</td>
<td>16</td>
<td>2,911</td>
</tr>
<tr>
<td>Further investigation not in the public interest</td>
<td>N/A1</td>
<td>1</td>
<td>153</td>
</tr>
<tr>
<td>Investigation complete; no suspect identified</td>
<td>Not cleared</td>
<td>8,477</td>
<td>1,233,654</td>
</tr>
<tr>
<td>Local resolution</td>
<td>Cleared</td>
<td>8</td>
<td>3,586</td>
</tr>
<tr>
<td>Offender deprived of property</td>
<td>Cleared</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Offender fined</td>
<td>Cleared</td>
<td>10</td>
<td>1,013</td>
</tr>
<tr>
<td>Offender given a caution</td>
<td>Cleared</td>
<td>41</td>
<td>5,216</td>
</tr>
<tr>
<td>Offender given a drugs possession warning</td>
<td>Cleared</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>Offender given absolute discharge</td>
<td>Cleared</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>Offender given community sentence</td>
<td>Cleared</td>
<td>57</td>
<td>9,599</td>
</tr>
<tr>
<td>Offender given conditional discharge</td>
<td>Cleared</td>
<td>2</td>
<td>1,900</td>
</tr>
<tr>
<td>Offender given penalty notice</td>
<td>Cleared</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>Offender given suspended prison sentence</td>
<td>Cleared</td>
<td>31</td>
<td>6,528</td>
</tr>
<tr>
<td>Offender ordered to pay compensation</td>
<td>Cleared</td>
<td>1</td>
<td>223</td>
</tr>
<tr>
<td>Offender otherwise dealt with</td>
<td>Cleared</td>
<td>2</td>
<td>1,164</td>
</tr>
<tr>
<td>Offender sent to prison</td>
<td>Cleared</td>
<td>235</td>
<td>32,242</td>
</tr>
<tr>
<td>Status update unavailable</td>
<td>N/A1</td>
<td>316</td>
<td>210,628</td>
</tr>
<tr>
<td>Suspect charged as part of another case</td>
<td>Cleared</td>
<td>75</td>
<td>18,200</td>
</tr>
<tr>
<td>Unable to prosecute suspect</td>
<td>Not cleared</td>
<td>551</td>
<td>57,471</td>
</tr>
<tr>
<td>Under investigation</td>
<td>N/A1</td>
<td>34</td>
<td>6,293</td>
</tr>
<tr>
<td>Unknown (Blank)</td>
<td>N/A1</td>
<td>2</td>
<td>828</td>
</tr>
</tbody>
</table>

1These are excluded as it is unclear from the category if an offender was identified and charged.
• Secondly, it could be expected that the percentage of dwellings that are detached may be negatively associated with crime clearance rates. This is based on that these types of dwellings are offset from their neighbouring dwellings and often the road serving them (i.e. by a garden or yard). This may prevent witnesses (neighbours or passersby) from either reporting the burglary which would allow the police to arrive more promptly (increasing the chance of apprehending the offender during the crime) or from being able to witness the burglar and provide a description which may help in their identification and apprehension.

• In a similar vein to how they may influence burglary rates, residential mobility and ethnic heterogeneity are also likely to negatively affect burglary clearance rates. This is because these factors are expected to regulate the ability of communities to establish shared values and social control (further information is available in Chapter 2). As such, suspected burglaries (in those communities) may be less likely to be investigated by their neighbours who could then provide a description of the burglar(s) or report the burglary while it is in progress.

• Also, although the number of households (within a LSOA) is justifiably used in the main analysis of this chapter, in this analysis the number of residents (in a LSOA) is used instead. This is because LSOAs with large amounts of people are more likely to provide greater anonymity for burglars. The number of residents (in a LSOA) is also likely correlated with the number of dwellings and so given the rationale for the former, the latter is omitted.

• A further two explanatory variables, police workload (burglaries per police officer) and distance to the nearest police station, are also included in this analysis. For the former, this is based on that greater workloads likely limit the resources available to investigate each crime. As such lower rates of burglary clearance are to be expected. For the latter, this is based on the premise that offences near a police station can be
responded to more quickly (if reported). It also requires fewer resources to visit and
investigate the crime location (as it is closer) so it is plausible that the police are more
likely to investigate it more thoroughly.

Also, although the distance between the offender’s home and the offence location is liable
to play some role in the likelihood of a burglary being cleared, it is not possible to include
this variable in this analysis. This is because for offences that are not cleared (where no
offender has been identified) it is impossible to know the distance that the (unknown)
offender travelled.

A summary of the variables included in this analysis including information about their
derivation is provided in Table 7.2. The results of the analysis (and the comparison
analyses) are presented in Table 7.1. The odds-ratios [OR] reported can be interpreted as
the multiplicative effect of a one-unit increase in the explanatory variable on the likelihood
of a burglary being cleared.
Dirty data in offence location preference analyses

In Chapter 7, the differences between the four approaches for analysing offence location choices were examined. That is by generating a series of synthetic offender and offence location datasets. Those datasets were then subjected to analyses from the four approaches and the number of incorrect results, in terms of type I, II and III errors, were compared. These analyses suggested that, all else being equal, that choice-based analyses are more likely to detect the effects of variables correctly than the other approaches. The data generating process used however ignored that the different types of analysis require different data and that these data can differ based on attrition. In this chapter the expected differences between the approaches and the findings from Chapter 7 are investigated further by incorporating the data attrition and the concept of dirty data into the data generating process. Techniques for reducing the number of errors in the statistical estimates in the two key (target-based and choice-based) approaches are also examined.
8.1. Data attrition

In Chapter 7, the analyses suggest that, all else being equal, that choice-based analyses are more likely to detect the effects of variables correctly than the other approaches. However, these hypotheses so far have been tested using equally sized datasets for each analysis. In reality, it is unlikely that equally sized datasets would be available for each type of analysis. To explain, the different types of analysis require different data and the different types of data differ in availability. To make things concrete, consider the attrition of (100) burglaries in the UK criminal justice system in Figure 8.1. Firstly, of the population of burglaries that occur, only around 60% are reported to the police (Office for National Statistics, 2017). Of these, around 89% are correctly recorded as burglaries (Her Majesty’s Inspectorate of Constabulary, 2014). As such, from the population of burglaries that occur in a given area, the police have the offence locations – which are required for target-based analyses - for around 50%. The other types of analyses (including choice-based analyses) require offender information and particularly the offenders’ (at-least approximate) home locations. These analyses, therefore, need cleared burglaries where the offender(s) are identified. Given the clearance rate for burglaries is often relatively small at about 10%\(^{23}\), only around 5% of the population of burglaries are typically available for these analyses.

Although the choice-based analyses favoured well against the target-based analyses (and the other types of analyses) in Chapter 7, the ten-fold decrease in the amount of data available in real world datasets will likely impact upon the statistical power of analyses conducted using this approach. This could be to the extent that despite the possible limitations of target-based analyses (see earlier), they could be better equipped for dealing

\(^{23}\) Calculated for the years 2004-2014 using police.uk data. See also Appendix A in Chapter 7.
with smaller samples of burglaries. Based on this, it is hypothesised that when creating a synthetic sample of burglaries and randomly sampling from this to mimic the attrition observed in real-world data (50% for target-based analyses, and 10% from those for choice-based analyses):

*Hypothesis 1:* When samples of burglaries are randomly drawn from a population of burglaries, at smaller population sizes the true effects of variables are more likely to be correctly detected as significant predictors of offending in target-based analyses than choice-based analyses.

*Hypothesis 2:* When samples of burglaries are randomly drawn from a population of burglaries, at larger population sizes the true effects of variables are to be

---

There are also other reasons for using target-based analyses over those that require cleared offences (i.e. choice-based analyses). One key reason is because of the sensitivity (and legality) of sharing offenders' home addresses to researchers, in many cases choice-based analyses will not be possible.
correctly detected as significant predictors of offending in choice-based analyses than target-based analyses.

While the above process assumes the likelihood of a burglary being cleared is random, research (e.g. Paré, Felson and Ouimet, 2007) suggests there are geographic variations in (burglary) clearance rates based on certain factors. If these factors are explanatory variables in the choice model or co-vary with those variables (e.g. in the case of the number of households and number of resident variables), the data used and the model’s estimates will be biased (see also Manski and Lerman (1977) for a technical description specifically in terms of the conditional logit model). More explicitly, these factors will cause burglaries in certain types of areas to be over- or under-represented in the data analysed by this approach. The result of this is that if they are more frequent, it will appear that the characteristics of those areas are more attractive than they should be and this will exaggerate the analysis’ estimated influence of that characteristic; and vice versa for those that are less frequent. Those variables will then be more or less likely to be correctly detected as predictors of offending. This can be compared to sampling bias where if some members of the population (of burglaries) are not equally likely to be included in the sample (that is analysed), the results of the analysis can be misrepresentative of the population. As such, based on incorporating the likelihood of a burglary being cleared (see Appendix A in Chapter 7) in the sampling of burglaries, it is expected:

_Hypothesis 3:_ When samples of _cleared_ burglaries are non-randomly drawn from a population of _reported and recorded_ burglaries, the true effects of variables are less likely to be correctly detected as significant positive predictors of offending in choice-based analyses than if they were randomly drawn.

_Hypothesis 4:_ When samples of _cleared_ burglaries are non-randomly drawn from a population of _reported and recorded_ burglaries, at smaller population sizes the true
effects of variables are more likely to be correctly detected as significant positive predictors of offending in target-based analyses than in choice-based analyses.

Hypothesis 5: When samples of cleared burglaries are non-randomly drawn from a population of reported and recorded burglaries, at larger population sizes the true effects of variables are more likely to be correctly detected as significant positive predictors of offending in choice-based analyses than in target-based analyses.

8.1.1. Methodology

Simulation process

To test these hypotheses, a similar simulation process (and general methodology) to that used in Chapter 7 and particularly for the comparisons of the target-based and choice-based analyses in hypotheses 1 and 2 in that chapter, is followed. More specifically, rather than generating datasets of 100, 250, 500, 750 and 1,000 burglaries which are then all included in the subsequent analyses, here population datasets are generated which contain 2,000, 5,000, 10,000, 15,000 and 20,000 burglaries. Samples of these burglaries are then drawn based on the rates of attrition described above to give datasets of reported and recorded burglaries containing 1,000, 2,500, 5,000, 7,500 burglaries which are then used in the target-based analyses. Lastly, samples from these datasets are drawn to give datasets of cleared burglaries containing 100, 250, 500, 750 and 1,000 burglaries that are used in the equivalent choice-based analyses.

For hypotheses 1, 2 and (half of the data in) 3 this sampling is random. Each burglary in the population (dataset) has an equal probability (50% based on the figures above) of being reported and recorded and therefore included in the target-based analyses. Each reported and recorded burglary then has an equal chance (10% based on the figures above) of being cleared and therefore included in the equivalent choice-based analyses. For half of the data used in hypothesis 3 and all of the data for hypotheses 4 and 5, the drawing of reported and
recorded burglaries from the population dataset remains the same. However, variations in clearance rates are incorporated when sampling for the choice-based analyses. Here, the likelihood of a burglary being cleared is determined by unequal probability (weighted random) sampling based on the results of an empirical analysis of the Buckinghamshire data.

For this analysis, the odds of a burglary being cleared is estimated using a multi-level Bernoulli regression of burglary clearance in the study area (see Appendix A in Chapter 7 for further information on these analyses). These odds are shown in Table 8.1 but to give an example, burglaries are 0.96 times as likely to be cleared (and included in the sample) for every additional 100 residents living within the LSOA where the burglary was committed. These variables are calculated as described in Chapter 7 (see Table 7.2).

**Analytical methodology**

For the previous hypotheses in Chapter 7, the measure of model performance was estimated in terms of the number of errors associated with a particular technique. For example, in the equivalent comparisons of the target-based and choice-based analyses presented for hypotheses 1 and 2 in Chapter 7, the hypotheses predicted specific results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated Odds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to nearest police station (km)</td>
<td>0.99</td>
</tr>
<tr>
<td>Ethnic heterogeneity (10%)</td>
<td>0.86</td>
</tr>
<tr>
<td>Percent of houses that are detached (10%)</td>
<td>1.04</td>
</tr>
<tr>
<td>Police workload (burglaries per police officer)</td>
<td>0.97</td>
</tr>
<tr>
<td>Poverty (£50,000)</td>
<td>1.09</td>
</tr>
<tr>
<td>Residential mobility (10%)</td>
<td>0.87</td>
</tr>
<tr>
<td>Residents (100)</td>
<td>0.96</td>
</tr>
</tbody>
</table>
regarding locational characteristics (that were estimated in both approaches) and compared the frequency of errors in the estimated effects of those variables. The observations, the presence or absence of an error for each locational characteristic variable, from one analysis could, therefore, be paired with the equivalent observation from the other analysis. The advantages of this matching and the use of the McNemar’s tests and YMO tests (e.g. statistical power) over alternatives could therefore be exploited. However, for all of these hypotheses except 3 which also uses the YMO test, these tests are not applicable. This is because these hypotheses concern the overall error rate across all (includable) variables to determine which type of analysis is least or most likely to detect incorrect effects. Therefore, because the choice-based analyses include and estimate the effects of proximity and this cannot be incorporated in target-based analyses, not all observations can be matched across analyses. As such, and because the data is still clustered (see earlier), Donner’s cluster-adjusted chi-square test (1989) is used based on its overall performance including those over other alternative tests (e.g. Jung, Ahn and Donner, 2001; Jeong, 2016). These tests were calculated using the ‘aod’ package (Lesnoff and Lancelot, 2006) in Stata 14 (StataCorp, 2015).

8.1.2. Results

Hypotheses 1 and 2 regard the correct detection of the effects of variables in target-based and choice-based analyses when the samples available for each analysis are randomly drawn from a population of burglaries. Overall, a significant Donner’s chi-square test (p < 0.01) indicates a statistically significant difference between the 33% of effects (across all variables and datasets) that were incorrectly detected in target-based analyses

---

25 An alternative to this that still incorporates the advantages of matched data is to use YMO tests and exclude the results regarding the distance variable. For comparison this approach was also tested but it gives the same pattern of results and so is not discussed further.
compared to 23% when the equivalent (random) sample is subjected to choice-based analyses.

Shown in Figure 8.2 are the percentage of errors (incorrectly detected effects) in the target-based and random sample choice-based (and non-random sample choice-based) analyses for each variable and for each population and effect size used to derive the data. Similar to the results in Chapter 7 for hypothesis 2 (which is an equivalent analysis but using equal and, for the target-based analyses, smaller samples), the proportion of errors varies across the variables for both analyses and tends to decrease with larger population and effect sizes (see earlier for likely explanations of this).

More specifically for hypotheses 1 and 2, the graph suggests and is supported by significant Donner’s chi-square tests (all $p < 0.01$) that errors are more likely in target-based analyses (red) than random sample choice-based analyses (blue) when the

---

**Figure 8.2: The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to target-based analyses and choice-based analyses using randomly sampled and non-randomly sampled cleared burglaries (hypotheses 1-5)**
population datasets contain 5,000 or more burglaries. This equates to 2,500 burglaries subjected to target-based analyses and 250 to choice-based analyses. When the population datasets contain 2,000 burglaries, which amounts to 1,000 burglaries subjected to target-based analyses and 100 to choice-based analyses, the difference (41% and 38% respectively) is not statistically significant ($p > 0.05$). In other words, at smaller population sizes (for 2,000 events at least) the number of errors do not significantly differ but at larger populations sizes choice-based analyses are more likely to detect the true effects of variables correctly. Also, as shown in Figure 8.2, Donner’s tests highlight that when the effect size of the variables is 1.1 the two approaches do not differ in their ability to detect the variables’ effects ($p > 0.05$), otherwise the choice-based analyses are more likely to correctly detect their effects (all $p < 0.05$).

In terms of hypothesis 3 and comparing random sample choice-based analyses and the more realistic non-random sample choice-based analyses, the significant YMO test indicates that the analyses using random sampling (which found errors in 23% of the variables across datasets) are overall less likely to find errors than those using non-random sampling (which found errors in 29% of the variables) ($p < 0.01$). As suggested in Figure 8.2, this result is also consistent across sample sizes (all $p < 0.05$). By looking at the blue and green bars in Figure 8.2, it also suggests that this pattern, where errors are more frequent for the non-random sample analyses, is consistent for three of the variables considered (ethnic heterogeneity, number of households and residential mobility); but that errors are more likely in random-sample choice-based analyses for the percent detached and poverty variables. This follows from what was expected when geographic variations in clearance rates are incorporated in the sampling. More specifically, burglaries in the types of areas which are less likely to be cleared (e.g. those with higher ethnic heterogeneity and residential mobility), will be less common and under-represented in the
datasets. As such, when analysed these variables will appear to be less attractive for burglaries and their (positive) effect will be more likely to go undetected or incorrectly detected. As shown in Figure 8.2, the reverse is true in for the percent detached and poverty variables as offences in these kinds of areas will be over-represented and appear as more attractive than they should.

Finally, for hypotheses 4 and 5, the correct detection of effects is compared between target-based analysis and the choice-based analyses using non-random sampling (to incorporate geographic variations in burglary clearance rates). Overall, the significant Donner’s test indicates the choice-based analyses are more likely to correctly detect the effects of the variables with errors found in 29% of the variables across datasets compared to 33% in the target-based analyses ($p < 0.05$).

As illustrated in Figure 8.2, when the population dataset contains 2,000 burglaries, and so 1,000 are available for the choice-based analysis and 100 for the target-based analysis, the difference between the two analyses (of 1%) is not significant ($p > 0.05$). For all larger sample sizes the non-random sample choice-based analyses outperform the target-based analyses in correctly detecting the effects of variables (all $p < 0.05$). When looking at the effect sizes used to generate the data, the pattern is more complicated. For three of the five effect sizes (1.05, 1.5 and 2) the choice-based analyses are less likely to incorrectly detect the effects of variables, however when the effect size is 1.1 or 1.25 the reverse is true (all $p < 0.05$).

### 8.2. Methods for improving target-based and choice-based estimates

So far this chapter has identified the limitations of the different approaches to analysing patterns of offence locations. This section will focus on testing the effects of
methodological corrections intended to address these limitations. Taking the two most promising approaches (as found in Chapter 7), the target-based and choice-based approaches, these limitations and possible corrections are as follows.

For target-based analyses, the critical limitation is the inability to include individual-alternative-specific variables, and in particular, proximity. This is because for this analytic approach independent variables can only vary over alternatives (they are alternative-specific variables). However, proximity will vary for each combination of individual (offender) and alternative (area), because the distance between each offenders’ home and each area will vary based on the location of the area and where each offender lives. Consequently, these analyses cannot natively account for the role of propinquity in offending patterns. One solution to this was employed by Bernasco and Luykx (2003) who used an aggregate measure of each area’s accessibility to (all) offenders called the \textit{spatially weighted burglar exposition rate (SWEBER)\textsuperscript{26}}. The aim of this approach is to take into account the role of proximity where an area is more vulnerable to offending (from all offenders) if it is closer to more offenders. This variable can be included in these types of (target-based) analyses as it only varies over alternatives (is alternative-specific). This SWEBER measure is calculated for these simulation analyses (see also the original in Bernasco and Luykx (2003)) by computing:

\begin{equation}
\text{SWEBER}_j = \sum_{i=1}^{N} M_i \times D_{ij}^{-2}
\end{equation}

\textsuperscript{26} A similar approach was also taken in Bernasco and Block (2011) which used the number of offenders living in each census block (their spatial unit). However, for these analyses the earlier method of SWEBER (Bernasco and Luykx, 2003) is preferred given it will also account, to some degree, for offenders living outside the focal spatial area but nearby.
Where, SWEBER for area $j$ is the sum of the number of burglaries committed by residents from area $i$ multiplied by the inverse square distance between areas $i$ and $j$. The choice of inverse square distance ($D_{ij}^{-2}$) as the functional form of distance decay follows Bernasco and Luykx (2003) and although this could be replaced by the known form in each analysis (as it is used to generate the data), doing so would be unrealistic since in real-life this would not be known.

For the simplest and (albeit unrealistic) evaluation of this measure, the home locations of all offenders that committed a reported and recorded burglary (those for whom the data were included and analysed in the target-based analysis in chapter X) could be used in the SWEBER measure:

_Hypothesis 6:_ Target-based analyses that incorporate the SWEBER measure using the home locations from all reported and recorded burglaries are more likely to detect the true effects of variables than the equivalent target-based analyses that ignore the role of offender accessibility.

That said, as described earlier, offenders’ home locations are only known for those who committed cleared offences. As such, in real-life applications, the SWEBER measure is estimated using only a subset of the offenders’ homes: those associated with cleared burglaries. Therefore to realistically assess the utility of including the SWEBER variable, to make the analyses more realistic and to account for the fact that cleared burglaries are a non-random sample of all burglaries, geographic variations in clearance rates can also be incorporated into the sampling approach when calculating the SWEBER measure:

27 Although the SWEBER measure could be weighted according to the probability an offence is cleared and the offender is known (see below for this being applied to choice-based analyses), it no discernible impact on the results so is omitted from what follows.
Hypothesis 7: When samples of cleared burglaries are non-randomly drawn from a population of reported and recorded burglaries, target-based analyses that incorporate the SWEBER measure using the home locations from cleared burglaries are more likely to detect the true effects of variables than the equivalent target-based analyses that ignore the role of proximity.

As explained earlier, the key limitation of the choice-based analyses is that the results are only as representative (of the wider population of offenders) as the data analysed. And, as clearance rates for burglaries are known to vary according to some factors (see earlier; also see Appendix A in Chapter 7) this can bias the results of the analyses. However, as the mechanism for this problem can be explained in terms of the non-random sampling of cleared burglaries from those that are reported and recorded, there are various treatments within the general literature that could be applied to address these issues. Here, two approaches are compared.

The first and the simplest is through sample (probability) weighting (see also Bernasco, Block and Ruiter, 2013 who also use the same approach). For a burglary spatial discrete choice model this would mean that each observation, a burglar’s selection of offence location, is weighted by the normalised (so the number of observations remain the same) reciprocal of the probability it would be included (the burglary being cleared based on the burglary location) in the sample analysed. Observations that are more likely to be included (e.g. those committed in areas with higher clearance rates) would therefore receive a smaller weight and so contribute relatively less to the model than observations which are less likely to be included. This weighting gives the Manski and Lerman (1977; see also Ben-Akiva and Lerman, 1985) weighted exogenous sample maximum likelihood (WESML) estimator of the conditional logit model (rather than the maximum likelihood estimator commonly used) which also uses robust standard errors (White, 1982) to adjust for
heteroscedasticity where clusters of observations contribute unequally (due to their weighting) to the model. Based on this it is expected that:

**Hypothesis 8:** When samples of *cleared* burglaries are non-randomly drawn from a population of *reported and recorded* burglaries, choice based-analyses calculated using WESML are more likely to detect the true effects of variables than the equivalent analysis using ML.

A second approach is through weighted bootstrap resampling. In a typical application of this technique, bootstrap samples are randomly and repeatedly taken, *with replacement*, from the (observed) sample of data. Parameter(s) are then estimated for each bootstrap sample and combined to form the standard error of the parameter estimate(s). In this application, this process is modified following a procedure such as that outlined in Nahorniak *et al.* (2015) whereby the probability of each observation being bootstrap sampled is the inverse of its probability of being included in the (observed) sample. To give an example, burglaries in less ethnically diverse areas, which are more likely to be cleared and so included in the observed sample, are weighted so that they are individually less likely to be included in each bootstrap sample. In effect, this process should approximate a random sample of cleared burglaries using the non-random sample of observed cleared burglaries (if there is sufficient information on clearance rates). Although typical bootstrap applications are only employed to estimate the variation of the bootstrapped estimates, the mean of those estimates can also be taken (as is done here) as the parameter estimate. One advantage of this bootstrapping is that because the non-randomness is removed (through resampling) before analysis, no weighting or robust standard errors need to be computed. This leads to:

**Hypothesis 9:** When samples of *cleared* burglaries are non-randomly drawn from a population of *reported and recorded* burglaries, choice based-analyses calculated using
weighted bootstrap resampling are more likely to detect the true effects of variables than the equivalent choice-based analysis.

Overall, although modifications are offered for target-based and choice-based analyses, it is still expected that because choice-based analyses naively incorporate all types of variables, when comparing the best performing realistic models (e.g. those in hypotheses 7):

**Hypothesis 10:** Choice-based analyses computed using either WESML or bootstrap resampling are more likely to correctly detect the effects of variables than the equivalent target-based analysis using SWEBER.

### 8.2.1. Methodology

**Simulation process**

To test hypotheses, a similar simulation process (and general methodology) to that used in 8.1 is followed. Notably, population datasets are generated to contain 2,000, 5,000, 10,000, 15,000 and 20,000 burglaries. 50% of these are then randomly sampled to give the datasets of reported and recorded burglaries (which are then subjected to target-based analyses). For all but hypothesis 15 (which incorporates all reported and recorded burglaries in the SWEBER measure), 10% of these burglaries are then non-randomly sampled (using the likelihood of a burglary being cleared from the regression model, see earlier and Appendix A in Chapter 7) to give the datasets of cleared burglaries (that are then subjected to choice-based analyses).

**Analytical methodology**

As introduced for this set of analyses, the SWEBER measures are calculated as described in equation 18. For the sampling weights, the standard calculation is used. More
specifically, using data on cleared and uncleared burglaries from police.uk\textsuperscript{28}, the reciprocal of the burglary clearance rates in each LSOA (the spatial units in these analyses) is normalised (so that the weighted number of observations equals the original sample size and not the population size) by multiplying it by the fraction of all burglaries in the analysed synthetic datasets that are cleared. Regarding the bootstrap methodology, due to the computational demands, 200 resamples (which are then combined) are taken from each generated dataset\textsuperscript{29}.

Finally, all hypotheses are tested using Yang modified-Obuchowski tests. For the majority of hypotheses (e.g. those comparing the same types of analyses) this is because the analyses include the same variables and these variables are measured in the same way. In the case of comparing the target-based and choice-based analyses (hypothesis 19), this test is also used because despite the target-based analyses incorporating the effect of accessibility using the SWEBER measures (of proximity across all offenders) and choice-based analyses using the proximity to each area for each offender, the interpretations of both variables are equivalent. For example, and considering that these hypotheses are only considering the frequency of type I, II and III errors in each type of analysis, if either measure is found to be non-significant then it is inferred that proximity does not play a role in offending for the burglars.

\textsuperscript{28} This data is used rather than the results from the Bernoulli analysis of burglary clearances (see Appendix A in Chapter 7) (which are used to generate the data) to purposefully create additional noise in the data to make it more realistic. For one, in real-world analyses the determinates of clearance rates would only be approximately known (e.g. through a similar analysis as in Appendix A in Chapter 7).

\textsuperscript{29} While 200 resamples is a relatively low number (e.g. many analyses use 1,000 or more), increasing this (e.g. to 500) had little effect on the overall results. Moreover, while increasing this is possible for real-life analyses, in these analyses this is not computationally feasible. This is because the 200 resamples (which are each subject to a choice-based analysis) are taken from every simulated dataset and there are 2,500 simulated datasets (100 for each combination of the five effect sizes and five population sizes).
8.2.2. Results

Hypotheses 6 and 7 regard the correct detection of the effects of variables in target-based analyses that incorporate or do not incorporate the SWEBER measure of overall proximity. As suggested in Figure 8.3, when SWEBER is calculated using the data generated for all synthetic offenders (not just those associated with cleared burglaries) the effects of all variables are significantly more likely to be detected than the equivalent target-based analyses without the SWEBER measure ($p < 0.01$). In fact, while the effects of 33% of the variables (across all datasets) are incorrectly detected using standard target-based analyses, this is true for only 12% of the variables in the SWEBER target-based analyses.

As shown in Figure 8.3 this large difference can be attributed to the substantial increase in the inability of the analytic approach detect the true effects of ethnic heterogeneity (down from 96% to 1%) and the modest increase in the residential mobility variable (24%)

**Figure 8.3:** The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to three versions of target-based analyses (hypotheses 6 and 7)
to 6%). That said, poverty is more likely to be incorrectly detected (as either not significant or significant but in the opposite direction) in SWEBER target-based analyses (37%) than in the equivalent non-SWEBER analyses (19%). The improvements associated with the SWEBER analyses is found regardless of the population sizes of the datasets and the magnitude of the effect sizes used to generate those datasets.

Even after adjusting for the fact that not all burglaries are cleared - by calculating the SWEBER measure using a non-random sample of burglars-, the SWEBER analyses still outperform the equivalent target-based analyses without the SWEBER measure ($p < 0.01$). In fact, the results from these SWEBER analyses (using a non-random sample of burglars to derive SWEBER) are very similar to those from the equivalent analyses where SWEBER is calculated using all (cleared and not cleared) burglaries (see also Figure 8.3) and the number of errors do not significantly differ (both around 12%, $p > 0.05$).

Hypotheses 8 and 9 both regard the correct detection of the effects of variables in choice-based analyses where those analyses are calculated as normal, using (sample weighting) WESML and using (unequal probability) bootstrap resampling. As suggested in Figure 8.4, in choice-based analyses using WESML, the effects of the variables are not significantly more (or less) likely to be correctly detected than the equivalent standard choice-based analysis ($p > 0.05$). In fact, overall errors are slightly more likely to be found for WESML choice-based analyses (30% compared to 29%). As shown in Figure 8.4, this is largely due to its loss of efficiency, likely due to the necessarily inflated robust standard errors, when dealing with smaller samples (population datasets containing 2,000 and 5,000 burglaries; i.e. analyses conducted on 500 and 1,000 burglaries) and when the datasets are generated with small effect sizes of the variables (1.05).
In comparison, choice-based analyses conducted using bootstrap resampling are significantly more likely to correctly detect the effects of variables than the equivalent standard choice-based analysis ($p < 0.01$). Overall, these analyses are unable to detect the effects of 25% of the variables (across datasets) compared to 29% of the variables in the standard choice-based analysis. This pattern is consistent across population dataset sizes and relatively consistent across the different effect sizes of variables used to generate the datasets.

When comparing the best performing (in terms of correctly detecting the effects of variables) target-based analyses (using SWEBER calculated from cleared burglaries) and choice-based analyses (using bootstrap resampling) for hypothesis 10, the target-based analyses are significantly (and substantially) more likely to correctly detect the effects of variables ($p < 0.01$). That is, incorrect effects were detected in only 12% of the variables (across all datasets) in the target-based analyses compared to 25% in the choice-based analyses.
analyses. As shown in Figure 8.5, this difference is relatively consistent (and still substantial) across each of the variables (in all but the poverty variable) and when the datasets are generated with different population sizes and effect sizes (see also Figures 8.6 and 8.7). That said, when the datasets are generated where the effect size of each variable is 1.5 or 2 the difference is negligible and not significant ($p > 0.05$).

### 8.3. Discussion

The aim of this chapter was to build on the analyses conducted in Chapter 7. In particular, to investigate the impact of (non-random) data attrition on the results from the target-based and choice-based approaches to analysing offence location choices. This included taking account of the fact that larger samples of data are typically available for target-based analyses than for choice-based analyses because the latter require data for detected offences, for which clear up rates are about one-tenth of those reported and recorded. Also modelled were the potential effects of detection bias, whereby there is non-random

---

**Figure 8.5: The percentage of variables that go undetected as significant positive predictors of offending by variable, population size and effect size according to the best performing target-based and choice-based analyses (hypothesis 10)**

- 193 -
data attrition in the clearance of offences according to factors that vary at the area level. While the results from these analyses generally supported the hypotheses and choice-based analyses were still more likely to correctly detect the effects of the variables, the differences between the two approaches became less substantial than for the first set of comparisons (but still statistically significant) (see also Figure 8.6).

Also, this chapter sought to examine techniques for improving the results and reducing the number of errors in the statistical estimates in the two approaches. In particular, three methods for addressing the shortcomings of the target-based and choice-based analyses were tested. For the target-based analyses this involved incorporating the otherwise omitted role of proximity using the SWEBER measure of overall offender accessibility. For the choice-based analyses, the potential effects of non-random variations in clearance rates were addressed through sample weighting (WESML estimation) and weighted bootstrap resampling. Although the results from these analyses supported hypotheses 6-9 and demonstrated that these methods improved the performance of each approach in correctly detecting the effects of variables known to be involved in the data generating process, contrary to hypothesis 10 the target-based analyses that incorporated the

Figure 8.6: The percentage of variables that go undetected as significant positive predictors of offending when using different clearance rates of datasets containing 1,000 reported and recorded burglaries according to the tested target-based and choice-based analyses
SWEBER measure outperformed the equivalent choice-based analyses. More specifically, as shown in Figure 8.6, when 1,000 burglaries are subjected to SWEBER (and non-SWEBER) target-based analyses, the best performing (bootstrap) choice-based analyses only becomes similarly adept when using around 50% of the same data; which is much higher than the amount that would be available (10%) given typical burglary clearance rates.

Taken together with results from Chapter 7, the results from these analyses suggest the choice-based approach to analysing offence location patterns is, as expected, theoretically superior to the alternative approaches. Particularly compared to the offender-based and mobility-based approaches (when analysing subgroups) which similarly require data on cleared offences. However, in practice, target-based analyses can be as accurate, or more so, in detecting the true effects of variables when considering the differences in the data required by (and empirically available for) the two approaches and when incorporating some aspect of the otherwise omitted effect of proximity in the analytical model.

Although not intended as such, and not described until now, the results also speak to the sensitivity of the approaches, notably the target-based and choice-based approaches, to the correct detection of the effects of variables on offence location choices. For example, in the final analyses (in 8.2) show that overall the target-based and choice-based approaches can correctly detect the effects of 33% and 29% of the variables respectively and these decrease to 25% and 12% using the bootstrap and SWEBER versions of the analyses. While these figures appear relatively high in that 1 in 4 effects are not correctly detected in the best performing choice-based analyses, these are the average percentage from analyses of data generated using a range of effect and sample sizes. As shown in Figure 8.7 that also uses the results from the third set of analyses, the bootstrap (and WESML) choice-based and SWEBER target-based analyses perform relatively well (e.g.
incorrect effects are detected in around 5% or fewer of the variables) when analysing datasets with more than around 500 burglaries and with expected effect sizes of 1.25 or greater.

These findings have clear implications for future research and practical applications of such research. The most obvious is in implicating which approach should be used and under what circumstances. For example, as shown in the decision tree in Figure 8.8, if offender home locations are unavailable and cannot be approximated then a target-based analysis can only be performed. If, however the home locations are unknown but can be approximated, a SWEBER target-based analysis is advised. In demonstrating the shortcomings of each approach, this chapter also highlights future avenues of research. For example, in terms of investigating data attrition and the potential impacts this may have on analyses. One practical application of these findings is with regards to journey-
Figure 8.8: A decision tree for selecting the approach and model for analysing offence location choices

Note: Although shown as a simple process there are cases (particularly those highlighted by a grey dashed line) where more than one type of analysis can be suggested. Taking those highlighted from left to right: the mobility-based and choice-based analyses should give equivalent results so either should be appropriate; in cases where the conditions are nearly met or other individual-alternative-specific variables are to be explored (e.g. idiosyncratic betweenness, see Chapter 4) then a choice-based analysis may be preferred; and if the home locations of all offenders cannot be confidently approximated or this assumption is too stringent, a standard target-based analysis should be used.
to-crime analyses and their use for predicting the likely home locations or anchor points of serial offenders. That is, where

the current protocol advocates the incorporation of a buffer space (e.g. Levine, 2015). However, and notwithstanding the general lack of supporting evidence of buffer spaces, this chapter suggests otherwise as research in support of offending buffer spaces (because as far as the author is aware the only analyses have been conducted into this use the offender-based approach) are disputable.

All that being said, the results in Chapter 7 and this chapter are not without limitations. Firstly, the synthetic data were generated using relatively simplistic offence location choice processes. In the first set of analyses particularly, many of the data used to test the hypotheses were generated where the offenders were simulated to have preferences for proximity or locational characteristics. In reality (e.g. Bennett, Wright and Wright, 1984; Wright and Decker, 1996 for burglars), it is to be expected that both factors influence offender decision-making. That said, these simplifications were necessary to demonstrate in what ways the results from the different approaches could be biased. Beyond this, in all of the data a still somewhat simplified offence location choice process was followed. This included ignoring processes likely to occur in real-life such as where offenders may provisionally select offence locations but they may

offend elsewhere as they find better targets during their crime trip. The simulated data therefore diverge to some degree from that expected in reality and this should be explored further. Similarly, the extent to which these results, which are based on synthetic data albeit using a real-world geographic area and real offender home locations, can be generalised is unknown. For example, if the offenders resided more uniformly or randomly through the study area then the measures of overall accessibility (such as SWEBER) would vary less (or not at all) between possible target locations. It would
therefore be unable to discern between the locations selected for offending and those not and so would not account for the effect of proximity and would fail to improve the results from the target-based analyses to that achieved by the choice-based analyses (see also Guimaraes, Figueirdo and Woodward, 2003). Replication is therefore needed to establish the prevalence of the effects found in these analyses.

Future replications should also consider the actual analyses performed. For one, across all analyses incorrect results were determined from being either type I, II or III errors. However, further analyses may consider that type III errors and detecting a significant relationship in an incorrect direction is more serious than the other types of errors. Errors of magnitude where the estimated effect is significantly different to that used to generate the data may also be investigated. Some of the hypotheses-specific techniques employed may also warrant further research. For example, the negative binomial regression model as used in the offender-based and mobility-based analyses are only one model that is arguably relevant for these analyses. The methods for detecting preferences for proximity (using Spearman’s correlation coefficients) and buffer spaces (using exact chi-square tests) in the offender-based analyses were also atypical. In addition, in choice-based analyses, preferences for proximity is usually determined by including distance (or proximity) as a single continuous predictor variable rather than as a series of dummy variables and computing the predicted frequencies in each dummy variable (as done here). Finally, and for the latter analyses, there are alternative techniques that may yield more accurate bootstrap confidence intervals, such as bias-corrected and accelerated intervals (Efron and Tibshirani, 1994). While many of these choices were justified, for example in ensuring the analyses were directly comparable, their effect on the overall results, including evaluating if these alternative techniques should be followed in future research, would benefit from further investigation.
Chapter 9: Systematic review and meta-analysis of offence location choice research

9

Systematic review and meta-analysis of offence location choice research

In the previous three chapters, the spatial discrete choice approach to analysing offence location patterns was introduced (Chapter 6) and assessed relative to the other alternative approaches (Chapters 7 and 8). In this chapter, a systematic review of this literature is conducted and the results from the previous applications of this approach (to offence location choices) are synthesised in a series of meta-analyses and meta-regressions. This chapter is motivated by the desire to obtain more accurate and reliable estimates of the factors that influence offence (including burglary) location choices and if and how those estimates are affected by different model specifications. The main contributions of this chapter are: 1) ascertaining which variables and model parameters should be used in future analyses (e.g. of burglaries in Chapters 10 and 11).

9.1. Introduction

As discussed in previous chapters, since their introduction in the now-seminal article “How do residential burglars select target areas?” by Wim Bernasco and Paul Nieuwbeerta in 2005, spatial discrete choice modelling has become a relatively popular technique for
analysing offending patterns. The related published literature now comprises of studies
that cover a variety of crime types from burglary (e.g. Bernasco and Nieuwbeerta, 2005)
and robbery (e.g. Bernasco and Block, 2009) to violence (Summers, 2012) and rioting
(Baudains, Braithwaite and Johnson, 2013). The analytical approach has also been applied
to a range of study areas including in the Netherlands (e.g. Bernasco and Nieuwbeerta,
2005) and England (e.g. Baudains, Braithwaite and Johnson, 2013) to the USA (e.g.
Bernasco and Block, 2009) and Australia (e.g. Clare, Fernandez and Morgan, 2009).

That said, despite the growing literature, there has been little formal collation and
synthesis of the research (though see Ruiter, 2017 for a general literature review). This is
important following the findings presented in Chapters 7 and 8 that show discrete choice
analyses that use smaller sample sizes may lead to errors of statistical inference. As such,
individual studies may be too small and limited to draw unequivocal conclusions regarding
the role of offence location choice criteria. In comparison, combining the results obtained
across studies in a formal meta-analyses can increase statistical power and provide more
accurate and reliable findings that are generalizable across the (included) population of
analyses.

Such a synthesis would also allow for the identification of any inconsistencies between
analyses which could identify challenges to theory or as new research questions to be
explained. For example, with the exception of Townsley et al.’s (2015a) analyses of
burglary target selections across three countries, there are no formal comparisons of the
factors that influence offender location choice across study areas. Also, excluding the
analyses in Bernasco (2010), which compared the parameter estimates for variables
specifically related to residential history across four types of crimes, there are no formal
comparisons of the estimated influences of offending between crime types. Therefore, at
present, the degree to which the (average) offender for one type of crime, or in one study
area differs to another is relatively unknown. This gap in knowledge is also especially relevant given the recent results in Townsley et al. (2015b) and Frith et al. (Frith, Johnson and Fry, 2017) which show preference variation between offenders for the same crime type in the same study area. Furthermore, in a meta-regression, moderators can be included to account for and estimate the influence of analysis-level differences (such as the inclusion or omission of variables) on the analysis results. Findings of which can be used to assess the validity of current and future discrete choice analyses of offence location choices.

Providing this information will not only inform future applications of this approach (such as that presented in Chapters 10 and 11), but it will also provide a greater basis for appraising those and previous analyses (e.g. in terms of the variables that should and were or were not included). The aim of this chapter is to thus conduct a systematic review of the offender spatial discrete choice literature. The results from relevant analyses are then statistically combined in a meta-analysis and meta-regression to derive a pooled overall estimate of the factors that influence offence location choices. Variation between studies will be analysed with a particular interest in explaining any variation, for example, in terms of the variables used, the study area analysed and other features (e.g. the spatial unit used) of each study.

The structure of the remainder of this chapter is as follows. In the next section, the systematic literature review process and results are described. In section 7.3, the meta-analysis and meta-regression methodologies are outlined. Next, in sections 7.4 and 7.5, the results from the meta-analyses and meta-regressions are presented. Finally, in the last section, 7.6, the findings from this chapter are discussed.
9.2. Systematic literature review

9.2.1. Search strategy

To identify relevant studies and analyses, the following search strategy was employed in October 2017. Firstly, all combinations of two sets of key-terms were used to search the titles, abstracts and keywords of entries in six electronic databases\(^{30}\). These terms were generated from the already-known general (e.g. Train, 2009) and crime-related (e.g. Bernasco and Nieuwbeerta, 2005) literature. The first set of terms related to the modelling approach and are *conditional logit*\(^{31}\), *discrete choice*, *discrete spatial choice*, *qualitative choice* and *random utility*. The second set of terms are related to their application to the crime domain and are *crime/criminal* and *offence/offender/offending*. The databases searched included one specifically related to crime and criminology, the *ProQuest Criminal Justice Database*; and four other, more general but popular, databases: *PubMed Central*, *ScienceDirect*, *Scopus* and *Web of Science*. In addition, to identify other studies that may be part of the grey literature (e.g. unpublished analyses) the *ProQuest Dissertations & Theses Global* database was also searched.

This search strategy was supplemented through backward and forward reference searching. More specifically, backward searches were conducted where relevant references found within the already identified papers are then reviewed. The forward search was also carried out where all papers citing the seminal paper, “How do residential burglars select target areas?” by Bernasco and Nieuwbeerta in 2005, were reviewed. The forward search

---

\(^{30}\) Other databases were also considered but they were either unavailable to the author (EBSCO Publishing’s Criminal Justice Abstracts database), not accessible (*National Criminal Justice Reference Service Abstracts Database*) or they lacked functional advanced search features (Mendeley, SAGE, SpringerLink, UCL Discovery and Wiley Online).

\(^{31}\) These search strategies were also repeated but including the term *multinomial logit* (which is a model similar to the conditional logit but strictly speaking excludes alternative-specific variables). This however only increased the number of non-relevant studies and so is not discussed further.
was conducted using two databases, *Scopus* and *Web of Science*, which gave different numbers (138 and 109 respectively) of entries citing this article.

Lastly, and partly to include other grey literature, other studies such as unpublished manuscripts (e.g. theses from the Department of Security and Crime Science at UCL) or those recently published (and so may not be indexed by those databases) that are known to the author were also included.

### 9.2.2. Inclusion criteria

From the studies identified from the search strategy outlined above, the study (and their analyses) were included in the next stage of the review if the following inclusion criteria were met:

1. The study must be written in English;
2. The study must contain a spatial discrete choice analysis, such as a conditional logit (CL) model, with at least one measure of proximity (to the offenders’ homes) and one regarding the attractiveness of alternatives. These criteria were used given their joint role in offence locations and because the key advantage of discrete choice methods is their ability to incorporate both types of variables in the same model.
3. The analysis must be related to crime or any related illegal behaviour, for example, anti-social behaviour as defined in the UK as behaviour “…that caused or was likely to cause harassment, alarm or distress to one or more persons…” (The Crime and Disorder Act 1998);
4. The (spatial) analysis must be of offence location choices.

### 9.2.3. Search results (studies)

Using this search strategy, 571 studies were retrieved. After removing 350 duplicates, the remaining 221 studies were evaluated using the four inclusion criteria above. Based on
this, 26 studies authored between 2005 and October 2017 were identified as relevant (see also Table 1). A breakdown of this process, including the number excluded using each of the criteria, is shown in the Sankey diagram in Figure 9.1. Of note, the potentially unexpected number of studies excluded using the fourth inclusion criteria (the analysis must be of offence location choice) is due to studies analysing a discrete spatial choice of non-offending behaviours, such as residential choice, that incorporate some measure of crime as an independent variable. Also, of the other studies excluded, Bernasco and Jacques (2015) and Smith and Brown (2007) warrant further comment. This is because despite analysing offence location choices, they did not include a proximity (to the offenders’ homes) variable and so were excluded based on the second inclusion criteria (see also earlier). Lastly, as shown in Figure 9.2, although many of the 26 studies were found through more than one element of the search strategy, no single element would have identified all studies and this is true even after excluding the five studies that were only identified because they were known to the author.

Figure 9.1: Sankey diagram showing the breakdown of studies identified and excluded based on each of the criteria
Using the inclusion criteria above to identify papers and then the individual analyses reported within those studies, 144 unique discrete choice analyses were found and extracted. Note that the unit of analyses are discrete choice analyses rather than studies.
due to some studies reporting multiple analyses. Also, that the six analyses in Baudains (2015) and Baudains et al. (2016) and two conditional logit analyses in Frith (2014) and Frith et al. (2017) were reported in both publications respectively and so are only counted once for each unique analysis. Also that analyses were recomputed if they include variables but do not report their effect and the raw data is available. This equates to around 5.5 analyses per study. Many of the analyses are however not distinct in terms of the data used (see also Table 9.1). In some studies, the same data is used for all analyses. For example, the seven analyses in Chamberlain and Boggess (2016) use the same data to compare competing models. In other studies, such as Menting (2017), the same data is used altogether (in this case to analyse all types of offenders) and is also divided to analyse subsets (in this case to analyse the different types of offenders separately). The same or overlapping data may also be used in analyses from different publications. For example, the same offending data is used in Bernasco and Block (2009), Bernasco et al. (2013) and Bernasco et al. (2017). Descriptive information of the studies and the analyses found within the studies is shown in Table 9.1 and summarised in Table 9.2.

In terms of the analyses, as can be seen in Tables 9.1 and 9.2, the majority are for offenders in Europe (73% of the analyses) and in particular in England or the Netherlands (both 35% of the analyses). Of those analyses from the Netherlands, most are for offenders living in The Hague (44 of the 50 analyses) which includes analyses of offenders offending and living anywhere in The Hague (29 analyses) or only those offending and living in three municipalities (The Hague, Rijswijk and Leidschendam-Voorburg) in The Hague (15 analyses). Outside of Europe, 33 analyses used data from the US including 26 that used offending data from Chicago. Six analyses used offending data from Australia, including four that used data from Brisbane. The other analyses are for offenders in Belfast in Northern Ireland (three analyses) and East Flanders in Belgium (one analysis).
### Table 9.1: Summaries of the 26 studies and their containing analyses as identified through the systematic literature review

<table>
<thead>
<tr>
<th>Reference</th>
<th>Published*</th>
<th>Study area(s)**</th>
<th>Offence type</th>
<th>Offence subtype(s)</th>
<th>No. of offences</th>
<th>Dataset time period</th>
<th>Spatial unit***</th>
<th>Spatial unit information</th>
<th>Model info</th>
<th>No. of analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baudains 2015</td>
<td>N</td>
<td>London (England)</td>
<td>Acquisitive</td>
<td>Rioting-related offences</td>
<td>232-1,477</td>
<td>7th–9th Aug 2011</td>
<td>Lower super output area</td>
<td>0.33km², 1,715 residents</td>
<td>CL</td>
<td>6</td>
</tr>
<tr>
<td>Baudains <em>et al.</em> 2013</td>
<td>Y</td>
<td>London (England)</td>
<td>Acquisitive</td>
<td>Rioting-related offences</td>
<td>232-1,477</td>
<td>7th–9th Aug 2011</td>
<td>Lower super output area</td>
<td>0.33km², 1,715 residents</td>
<td>CL</td>
<td>3</td>
</tr>
<tr>
<td>Baudains <em>et al.</em> 2016</td>
<td>Y</td>
<td>London (England)</td>
<td>Acquisitive</td>
<td>Rioting-related offences</td>
<td>232-1,477</td>
<td>7th–9th Aug 2011</td>
<td>Lower super output area</td>
<td>0.33km², 1,715 residents</td>
<td>CL</td>
<td>6</td>
</tr>
<tr>
<td>Bernasco 2006</td>
<td>Y</td>
<td>The Hague (Netherlands)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>365-809</td>
<td>1996-2004</td>
<td>Neighbourhood</td>
<td>0.65km², 4,952 residents</td>
<td>CL</td>
<td>2</td>
</tr>
<tr>
<td>Bernasco 2010a</td>
<td>Y</td>
<td>The Hague**** (Netherlands)</td>
<td>Various</td>
<td>Various, assault, robbery, burglary</td>
<td>≤ 5,994</td>
<td>2004-2005</td>
<td>Four digit postcode area</td>
<td>10.37km², 4,900 residents</td>
<td>CL</td>
<td>6</td>
</tr>
<tr>
<td>Bernasco 2010b</td>
<td>Y</td>
<td>The Hague (Netherlands)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>1,871</td>
<td>2002-2007</td>
<td>Six digit postcode area</td>
<td>0.02km², 40 residents</td>
<td>CL</td>
<td>5</td>
</tr>
<tr>
<td>Bernasco and Block 2009</td>
<td>Y</td>
<td>Chicago (US)</td>
<td>Acquisitive</td>
<td>Robbery</td>
<td>5,847</td>
<td>1996-1998</td>
<td>Census tract</td>
<td>0.72km², 3,298 residents</td>
<td>CL</td>
<td>2</td>
</tr>
<tr>
<td>Bernasco and Kooistra, 2010</td>
<td>Y</td>
<td>Netherlands</td>
<td>Acquisitive</td>
<td>(Commercial) Robbery</td>
<td>276</td>
<td>2004-2005</td>
<td>Four-digit postcode area</td>
<td>10.37km², 4,900 residents</td>
<td>CL</td>
<td>6</td>
</tr>
<tr>
<td>Bernasco and Nieuwbeerta, 2005</td>
<td>Y</td>
<td>The Hague (Netherlands)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>548</td>
<td>1996-2001</td>
<td>Neighbourhood</td>
<td>0.65km², 4,952 residents</td>
<td>CL</td>
<td>2</td>
</tr>
<tr>
<td>Study</td>
<td>Year</td>
<td>Location</td>
<td>Methodology</td>
<td>Offence</td>
<td>N</td>
<td>Year Range</td>
<td>Area Size</td>
<td>Population</td>
<td>CL</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
<td>------------------------</td>
<td>---------------</td>
<td>---------</td>
<td>-------</td>
<td>---------------------</td>
<td>-----------------</td>
<td>-------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td>Bernasco et al. 2013</td>
<td>Y</td>
<td>Chicago (US)</td>
<td>Acquisitive</td>
<td>Robbery</td>
<td>12,938</td>
<td>1996-1998</td>
<td>0.02km², 118 residents</td>
<td>WESMLE</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Bernasco et al. 2015</td>
<td>Y</td>
<td>West Midlands (England)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>3,337</td>
<td>2009-2012</td>
<td>Lower super output area</td>
<td>0.51km², 1,500 residents</td>
<td>CL</td>
<td>1</td>
</tr>
<tr>
<td>Bernasco et al. 2017</td>
<td>Y</td>
<td>Chicago (US)</td>
<td>Acquisitive</td>
<td>Robbery</td>
<td>~200-1900</td>
<td>1996-1998</td>
<td>Census block</td>
<td>0.02km², 118 residents</td>
<td>CL</td>
<td>19</td>
</tr>
<tr>
<td>Chamberlain and Boggess 2016</td>
<td>Y</td>
<td>Tampa (US)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>5,182</td>
<td>2000-2012</td>
<td>Census block</td>
<td>1.36km², 909 residents</td>
<td>CL</td>
<td>7</td>
</tr>
<tr>
<td>Clare et al. 2009</td>
<td>Y</td>
<td>Perth (Australia)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>1,761</td>
<td>2001-2002</td>
<td>Residential suburb</td>
<td>6.76km², 4,669 residents</td>
<td>CL</td>
<td>2</td>
</tr>
<tr>
<td>Frith 2012</td>
<td>N</td>
<td>York (England)</td>
<td>Acquisitive</td>
<td>Various, burglary, burglary non-dwelling, robbery, TFV, TOV</td>
<td>130-1,242</td>
<td>2008-2012</td>
<td>Output area</td>
<td>0.44km², 293 residents</td>
<td>CL</td>
<td>16</td>
</tr>
<tr>
<td>Frith 2014</td>
<td>N</td>
<td>High Wycombe (England)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>459</td>
<td>2004-2014</td>
<td>Street segment</td>
<td>0.03km², 40 residents</td>
<td>CL / MSL ML</td>
<td>6</td>
</tr>
<tr>
<td>Frith et al. 2017</td>
<td>Y</td>
<td>High Wycombe (England)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>459</td>
<td>2004-2014</td>
<td>Street segment</td>
<td>0.03km², 40 residents</td>
<td>CL / HB ML</td>
<td>5</td>
</tr>
<tr>
<td>Johnson and Summers 2015</td>
<td>Y</td>
<td>Dorset (England)</td>
<td>Acquisitive</td>
<td>TFV</td>
<td>721</td>
<td>2001-2005</td>
<td>Lower super output area</td>
<td>13.40km², 1,524 residents</td>
<td>CL</td>
<td>2</td>
</tr>
<tr>
<td>Lammers et al. 2015</td>
<td>Y</td>
<td>The Hague (Netherlands)</td>
<td>Various</td>
<td>Various</td>
<td>12,639</td>
<td>2009</td>
<td>Four digit post code area</td>
<td>2.96km², 7,000 residents</td>
<td>CL</td>
<td>6</td>
</tr>
<tr>
<td>Marchment 2015</td>
<td>N</td>
<td>Belfast (Northern Ireland)</td>
<td>Non-acquisitive</td>
<td>Terrorism</td>
<td>148</td>
<td>1969-1989</td>
<td>Ward</td>
<td>2.74km², 5,994 residents</td>
<td>CL</td>
<td>3</td>
</tr>
<tr>
<td>---------------</td>
<td>----</td>
<td>---------------------------</td>
<td>-----------------</td>
<td>-----------</td>
<td>-----</td>
<td>------------</td>
<td>------</td>
<td>--------------------------</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>Menting 2017</td>
<td>Y</td>
<td>Three municipalities in The Hague (Netherlands)</td>
<td>Various</td>
<td>Various, Burglary, Violence</td>
<td>2,970-13,088</td>
<td>2009</td>
<td>Four-digit postcode area</td>
<td>1.68km², 7,700 residents</td>
<td>CL</td>
<td>15</td>
</tr>
<tr>
<td>Menting <em>et al.</em> 2016</td>
<td>Y</td>
<td>The Hague (Netherlands)</td>
<td>Various</td>
<td>Various</td>
<td>19,420</td>
<td>2006-2009</td>
<td>Four digit postcode area</td>
<td>2.96km², 7,000 residents</td>
<td>CL</td>
<td>6</td>
</tr>
<tr>
<td>Summers 2012</td>
<td>N</td>
<td>London (England)</td>
<td>Non-acquisitive</td>
<td>Violence</td>
<td>276</td>
<td>April 2002 – March 2007</td>
<td>Lower super output area</td>
<td>0.33km², 1,505 residents</td>
<td>CL</td>
<td>12</td>
</tr>
<tr>
<td>Townsley <em>et al.</em> 2015a</td>
<td>Y</td>
<td>Birmingham (England)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>291</td>
<td>2009</td>
<td>Middle super output area</td>
<td>2.04km², 8,585 residents</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Townsley <em>et al.</em> 2015b</td>
<td>Y</td>
<td>Brisbane (Australia)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>273</td>
<td>2006</td>
<td>Statistical local area</td>
<td>8.48km², 7,495 residents</td>
<td>CL</td>
<td>2</td>
</tr>
<tr>
<td>Townsley <em>et al.</em> 2015b</td>
<td>Y</td>
<td>The Hague (Netherlands)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>290</td>
<td>1996-2001</td>
<td>Neighbourhood</td>
<td>0.65km², 4,952 residents</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Vandeviver <em>et al.</em> 2015</td>
<td>Y</td>
<td>East Flanders (Belgium)</td>
<td>Acquisitive</td>
<td>Burglary</td>
<td>650</td>
<td>2006-2012</td>
<td>Dwelling</td>
<td>&lt;0.01km², 2 residents</td>
<td>CL</td>
<td>1</td>
</tr>
</tbody>
</table>

* Published refers to if the study was published in a peer-reviewed journal (Y) or not (N)

**Study area refers to the geographic location of the offending data

*** Spatial unit refers to the alternatives that offenders are choosing between

Note: TFV=Theft from vehicle, TOV=theft of vehicle, CL=(maximum likelihood estimated) conditional logit, WESMLE CL= weighted exogenous sample maximum likelihood estimated conditional logit, HB ML=hierarchical Bayes estimated mixed logit, MSL ML=maximum simulated likelihood estimated mixed logit
Table 9.2: Summary information of the unique (144) analyses and studies found through the literature review

<table>
<thead>
<tr>
<th>Study area</th>
<th>No. of analyses</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Europe</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belgium</td>
<td></td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>East Flanders</td>
<td></td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>England</td>
<td></td>
<td>51</td>
<td>35%</td>
</tr>
<tr>
<td>Birmingham</td>
<td></td>
<td>2</td>
<td>1%</td>
</tr>
<tr>
<td>Dorset</td>
<td></td>
<td>2</td>
<td>1%</td>
</tr>
<tr>
<td>High Wycombe</td>
<td></td>
<td>9</td>
<td>6%</td>
</tr>
<tr>
<td>London</td>
<td></td>
<td>21</td>
<td>15%</td>
</tr>
<tr>
<td>West Midlands</td>
<td></td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>York</td>
<td></td>
<td>16</td>
<td>11%</td>
</tr>
<tr>
<td>Netherlands</td>
<td></td>
<td>50</td>
<td>35%</td>
</tr>
<tr>
<td>All of Netherlands</td>
<td></td>
<td>6</td>
<td>4%</td>
</tr>
<tr>
<td>The Hague</td>
<td></td>
<td>44</td>
<td>31%</td>
</tr>
<tr>
<td>Northern Ireland</td>
<td></td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Belfast</td>
<td></td>
<td>33</td>
<td>23%</td>
</tr>
<tr>
<td><strong>North America</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US</td>
<td></td>
<td>33</td>
<td>23%</td>
</tr>
<tr>
<td>Chicago</td>
<td></td>
<td>26</td>
<td>18%</td>
</tr>
<tr>
<td>Tampa</td>
<td></td>
<td>7</td>
<td>5%</td>
</tr>
<tr>
<td><strong>Oceania</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td></td>
<td>6</td>
<td>4%</td>
</tr>
<tr>
<td>Brisbane</td>
<td></td>
<td>4</td>
<td>3%</td>
</tr>
<tr>
<td>Perth</td>
<td></td>
<td>2</td>
<td>1%</td>
</tr>
<tr>
<td><strong>Offence type</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Acquisitive</strong></td>
<td></td>
<td>104</td>
<td>72%</td>
</tr>
<tr>
<td>Burglary (domestic)</td>
<td></td>
<td>39</td>
<td>27%</td>
</tr>
<tr>
<td>Burglary (domestic other)</td>
<td></td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>Rioting</td>
<td></td>
<td>9</td>
<td>6%</td>
</tr>
<tr>
<td>Robbery (street)</td>
<td></td>
<td>29</td>
<td>20%</td>
</tr>
<tr>
<td>Robbery (commercial)</td>
<td></td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>Theft from vehicle (TFV)</td>
<td></td>
<td>6</td>
<td>4%</td>
</tr>
<tr>
<td>Theft of vehicle (TOV)</td>
<td></td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Various</td>
<td></td>
<td>11</td>
<td>8%</td>
</tr>
<tr>
<td><strong>Non-acquisitive</strong></td>
<td></td>
<td>21</td>
<td>15%</td>
</tr>
<tr>
<td>Terrorism</td>
<td></td>
<td>3</td>
<td>2%</td>
</tr>
<tr>
<td>Violence/Assault</td>
<td></td>
<td>18</td>
<td>13%</td>
</tr>
<tr>
<td><strong>Various</strong></td>
<td></td>
<td>19</td>
<td>13%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model type and estimation method</th>
<th>No. of analyses</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conditional logit (CL)</strong></td>
<td></td>
<td>137</td>
<td>95%</td>
</tr>
<tr>
<td>Maximum likelihood estimator (MLE)</td>
<td></td>
<td>132</td>
<td>92%</td>
</tr>
<tr>
<td>Weighted exogenous sample maximum likelihood estimator (WESMLE)</td>
<td>5</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td><strong>Mixed logit (ML)</strong></td>
<td></td>
<td>7</td>
<td>5%</td>
</tr>
<tr>
<td>Hierarchical Bayes (HB)</td>
<td></td>
<td>4</td>
<td>3%</td>
</tr>
<tr>
<td>Maximum simulated likelihood (MSL)</td>
<td></td>
<td>3</td>
<td>2%</td>
</tr>
</tbody>
</table>
The majority of the analyses were of acquisitive offences (72% of offences). This includes 39 analyses of (domestic) burglary and 29 analyses of (street) robbery offences (and an analysis of commercial robberies). There were also 21 analyses (15% of all analyses) of non-acquisitive offences including 18 analyses of violence or assault. The remaining 13% of analyses (18 analyses) were of a mix of (acquisitive and non-acquisitive) offences. All analyses of non-acquisitive and mixed types of offences used data from Europe and most used data from The Hague in the Netherlands (24 analyses).

Further, in almost all analyses (137 of the 144 analyses), the standard discrete choice model, the CL, was used. That said, in five of these analyses these were estimated using the weighted exogenous sample maximum likelihood estimator [WESMLE] rather than maximum likelihood estimator (MLE) which weights the offence data so that it appears to be a simple random sample of all offences (see also Chapter 8). In the remaining six analyses, the mixed logit [ML] model was used. In three of these ML analyses, all from Frith (2014), the models were estimated using maximum simulated likelihood [MSL] which can be problematic as the models include a large choice set (Hensher and Greene, 2003; Townsley et al., 2015b; Frith, Johnson and Fry, 2017). That said, two of those three analyses were re-estimated in Frith et al. (2017) using hierarchical Bayes (which does not suffer from those issues).

As illustrated in Figure 9.3, studies vary considerably in terms of the of the spatial units (in terms of how the continuous space is discretized into alternatives that can be chosen by an offender) used. In terms of their geographic size, they can be approximately divided into three or four groups. The smallest are micro-level units as used in the analysis of individual dwellings (1 analysis). Next are the street-centric units, such as street segments, Dutch full (six-digit) postcode areas and US census blocks, which are used in 38 (26%) of the analyses. Used in 32% of the analyses are relatively small-scale local neighbourhoods (e.g.
smaller than 1km²) such as UK LSOAs and Dutch neighbourhoods in urban areas. Lastly, a large number of the analyses (41% of the analyses) use larger scale neighbourhoods. These can also be divided into those that are relatively smaller (e.g. between 1 and 5km²) which includes 27% of the analyses, and those which use units such as US census block groups and Dutch partial (four-digit) postcode areas. Then there are 20 analyses (14% of all analyses) that use much larger spatial areas such as UK LSOAs and Dutch partial (four-digit) postcode areas in or including more rural areas. The problem with comparing studies on geographic size is that the geopolitical units often used are generally defined to
create units with similar numbers of residents and so urban density can massively affect their size (e.g. in terms of UK LSOAs which can vary from an average size of around 0.4km\(^2\) in London to 13.4km\(^2\) in Dorset). As such, and although perhaps less interpretable, they can also be aggregated by the number of residents. That said, this metric creates relatively similar groups: those with very few (~2) residents (the analysis of dwellings), ~50-300 residents (e.g. street segments, US census blocks and UK output areas), ~1000-2000 (US census block groups and UK lower super output areas) and those larger (such as Dutch partial postcodes and neighbourhoods).

Lastly, a large variety of variables were tested for their impact on offence location choices. Even when those variables broadly measuring the same concepts were grouped, considerably more than 100 variables were used in the analyses. The most common of these groups of variables are shown in Table 9.3 along with the number of analyses that used them and the number for which the effect size was shown. Also shown in Table 9.3 are the results from a basic (vote-count) summary of those variables calculated by counting the number of analyses for which each variable was found to be a significant positive predictor, a significant negative predictor or a non-significant predictor of offending. The results show that the effect of many of the variables appear clear-cut. For example, distance to the offender’s home was a negative predictor (i.e. they prefer those closer to their home) of location choice in 99% of the 95 analyses (where the effect size was shown). For other variables, the effect is relatively inconsistent. For example, affluence is a negative predictor of offending in 27% of the 52 analyses, a positive predictor in 17% and a non-significant predictor in 56%.
While this vote-count method used here can be useful for summarising the roles of the variables in offence location choices, this method is not without serious limitations. Firstly, vote counting treats all analyses as equal and so can mask overall effects due to variations in sample and effect sizes where small samples that find non-significant results (due to lower statistical power) receive as much weight as analyses based on large samples with significant and marginal or substantial effects. In the case of this literature, this approach is also particularly limited because it cannot incorporate the (complete and partial) overlap in samples across analyses (see also earlier) where the effect sizes from these analyses will be correlated. One consequence of this for vote counting is that the results will be biased towards those datasets (and study areas, spatial units etc.) analysed

<table>
<thead>
<tr>
<th>Variable</th>
<th>Num. of analyses</th>
<th>Most common effect</th>
<th>OR&lt;1</th>
<th>NS</th>
<th>OR&gt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance to the offender’s home</td>
<td>121 95</td>
<td>Negative</td>
<td>99%</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Number of targets</td>
<td>80 69</td>
<td>Positive</td>
<td>3%</td>
<td>23%</td>
<td>74%</td>
</tr>
<tr>
<td>Affluence</td>
<td>63 52</td>
<td>Not significant</td>
<td>27%</td>
<td>56%</td>
<td>17%</td>
</tr>
<tr>
<td>Residents of the same ethnicity as the offender</td>
<td>46 46</td>
<td>Positive</td>
<td>0%</td>
<td>24%</td>
<td>76%</td>
</tr>
<tr>
<td>Residential history of the offender</td>
<td>73 43</td>
<td>Positive</td>
<td>14%</td>
<td>0%</td>
<td>86%</td>
</tr>
<tr>
<td>Residential churn</td>
<td>44 43</td>
<td>Not significant</td>
<td>19%</td>
<td>51%</td>
<td>30%</td>
</tr>
<tr>
<td>Presence of main streets</td>
<td>43 43</td>
<td>Positive</td>
<td>0%</td>
<td>40%</td>
<td>60%</td>
</tr>
<tr>
<td>Distance to the city centre</td>
<td>42 41</td>
<td>Not significant</td>
<td>10%</td>
<td>56%</td>
<td>34%</td>
</tr>
<tr>
<td>Presence of train stations</td>
<td>40 40</td>
<td>Positive</td>
<td>0%</td>
<td>28%</td>
<td>73%</td>
</tr>
<tr>
<td>Presence of schools</td>
<td>61 37</td>
<td>Positive</td>
<td>5%</td>
<td>38%</td>
<td>57%</td>
</tr>
</tbody>
</table>

OR=odds ratio, NS=not significant.
on more occasions. Due to these weaknesses, more sophisticated formal meta-analyses and meta-regressions are conducted.

9.3. Meta-analysis and meta-regression methodology

In contrast to other methods such as vote-count summaries (as above), meta-analyses (including meta-regressions) empirically synthesise quantitative literatures by statistically combining the estimated effect sizes whilst taking account of the precision of those estimates. This allows for the estimation the mean effect of each variable and its reliability (in statistical terms), rather than merely counting whether how often it is statistically significant and the most common direction of its effect. Meta-analysis can also be used to (statistically) account for other factors that are problematic in vote-count summaries. For example, sample sizes and correlated effects (e.g. from repeated use of the same or related datasets). Going further, meta-regressions can also incorporate moderator variables to account for and determine the influence of analysis-level differences such as different (additional) predictor variables.

As already discussed in Chapter 4 in the meta-analysis of configurational methods, general recommendations are followed whereby the meta-analyses are calculated using random-effects (see Chapter 4).

While meta-analyses can be used to combine many types of statistics, they are not without limitations. Firstly, they assume that each of the estimates to be combined are sufficiently similar in the sense that they represent (at least approximately) the same thing. In that way, the combined effect can be meaningfully interpreted. In the case of this literature, the estimates are however from different choice models (CL and ML), using different CL (MSL and WESMLE) and ML (MSL and HB) estimators and from analyses of different offence types, study areas and using different spatial units (see also above). Furthermore,
combining estimates from multivariable analyses (such as discrete choice analyses) is especially difficult as the estimated effect of each variable is net of the estimated effects of the other variables; it is a partial effect. As such, the estimated effect of each variable will depend (to some degree) on what other variables are also included in that analysis. The strictest implication of this is that each effect can only be meta-analysed if it was calculated with the same additional predictor variables as all other models. In this (and most other) literatures, this will severely limit the number and breadth of estimates that can be pooled. In fact, variation in (the variables included in the) statistical models should be expected as subsequent studies should develop and elaborate from those previous.

Although certain assumptions (of equivalence) can resolve this issue, an alternative is to use a meta-regression model. Here, differences between analyses can be accounted for using fixed-effect analysis-level moderator variables\textsuperscript{32} (e.g. Becker and Wu, 2007; Jackson and Riley, 2014; van Houwelingen, Arends and Stijnen, 2002). Adding these covariates (to a random-effects model) means that the estimated effects of each variable still belong to the same population of effects but they can differ from the mean by some observed between-study difference (the covariate). When these covariates are fixed, they have a constant effect on the estimated effects of each variable. These covariates may also be hypothesis tested to determine the significance and severity of their effect on the estimated effects of other variables. Despite these advantages, due to the limited number of estimates for some variables and the requirement of a sufficient number of estimates per moderator variable, meta-regressions can only be applied to a small number of variables. Based on this, in what follows meta-analyses and meta-regressions will both be calculated and the above issues and how they are resolved are now discussed.

\textsuperscript{32} By adding fixed-effect moderator variables to a random-effect meta-analysis this notationally creates a mixed-effects meta-regression.
9.3.1. Different effect indices

The first potential issue is that the effect indices to be combined need to be sufficiently similar in that they represent (at least approximately) the same thing. In the case of these analyses, there are three concerns.

The first regards the use the ML model in some of the analyses whereas the majority use CL. As described in Chapter 6, the main difference between these is that the latter assumes all offenders have the same preference for each attribute (or that they only systematically differ based on observed and modelled offender characteristics) and estimates those as fixed preferences. In contrast, the ML allows for preference variation amongst the offenders and estimates this distribution (including its mean and standard deviation). In other words, it estimates the average preference for each attribute. Although these effect indices should not differ substantially, research has shown this is often not the case. For example, although Dahlberg and Eklöf (2003) and Persson (2002) found relatively similar estimates, Bhat (2000) and Revelt and Train (1998) found they can differ somewhat substantially. Differences can include the coefficients estimated using a CL model being smaller in magnitude than the mean coefficients estimated using a ML model. This occurs depending on the amount of preference variation and is due to the normalisation of the parameters such that the extreme value terms have the appropriate variance (see also Chapters 6 and Revelt and Train, 1998). In the case of the CL, the extreme value term incorporates all variance (including between decision-makers) whereas the ML treats it separately. The variance in the error term is therefore greater in the CL, and the normalisation will make the parameters smaller in magnitude. As shown in Dahlberg and Eklöf (2003), this is also likely exacerbated in more parsimonious models where heterogeneity that would otherwise be explained by other variables is not explained. Considering this, and the fact that the true amount of heterogeneity is unknown, though
Frith (2014), Frith et al. (Frith, Johnson and Fry, 2017) and Townsley et al. (2015b) suggests there could be significant amounts, both estimates are included in the meta-analysis. For the meta-regression, moderator variables are included for the type of model.

The second related concern regards the use of different estimators in the choice models in some analyses. This includes the use of the WESMLE in the CL analyses in Bernasco et al. (2013) whereas the other CL analyses use the MLE. As explained in Chapter 6, WESMLE attempts to correct for unequal probabilities of offences being cleared such that the point estimates are more accurate for the underlying population (rather than the sample) of that study area. The WESMLE also requires robust standard errors (White, 1982) which are typically larger and more conservative (e.g. Wooldridge, 2015). WESMLE can, therefore, have qualitatively different statistical properties. Nevertheless, given the results in Bernasco et al. (2013) which suggest the WESMLE estimates should not substantively differ from the MLE equivalents, both are included in the meta-analysis but a moderator variable is included for the type of estimator in the meta-regression.

The final concern relates to the use of the MSL estimator of the ML in Frith (2014) whereas the other ML analyses use hierarchical Bayes (Frith, Johnson and Fry, 2017; Townsley et al., 2015b). Here, however, because of the advantages of hierarchical Bayes [HB] (see also Hensher and Greene, 2003) and that Frith et al. (Frith, Johnson and Fry, 2017) uses HB and runs almost all the same analyses, the MSL estimated analyses in Frith (2014) can simply be ignored.

9.3.2. Different predictor variables

The second possible issue is that of the non-equivalence of the statistical models in terms of the predictor variables that are included. That is, the estimated effect of each variable depends on what other variables are also included in that analysis. An example of this is in Frith (2014) where individual variables and groups of variables are gradually added to
the model (to explain burglary location choices). Here, the odds-ratio effect of distance (to the possible offence location) changes from 0.30 when distance is entered on its own to 0.44 when entered alongside a measure of awareness space and 0.47 when entered alongside other variables such as those concerning social disorganisation.

Whilst these differences can have a large impact on the estimates, there are scenarios where differences in model specifications should not have a great impact. For example, if each of the models are (reasonably) well-specified then any changes in the predictor variables should not (greatly) affect the estimated effects of the variables. Alternatively, if the predictor variables that vary between models are relatively independent of the other variables then the estimates should also not be effected substantially. The extent to which these scenarios apply is relatively unclear. However, given that the sequential addition of explanatory variables can affect existing coefficient estimates (e.g. Frith, 2014) differences in model specifications should be considered.

As such, for the meta-analysis, when multiple models are based on the same data, the estimates are only included from the most specified model that fits the data the best (i.e. the model with the largest $r^2$ value, or in its absence, the model with the most variables included and therefore accounted for). For the meta-regression, moderator variables are included for the presence or absence of variables or groups of them. For example, for Frith (2014) binary covariates are added for specific variables, such as idiosyncratic betweenness (see Chapter 5), and related groups of variables, such as social disorganisation. Also, due to using binary moderators to account for the inclusion or not of a group of related variables in an analysis, some (incomplete) analyses of the same data that only include some of that group (of variables) are excluded. One example regards the moderator representing if analyses account for the effect of the offender’s residential history. Here, in some analyses, residential history is included using only one variable (e.g.
1/0 to indicate if an area contains the former residential area of the offender or not) while in others it is included using multiple variables (e.g. that also take into account the length of time they resided in an area and when they left that area). In this example, the former would be excluded if the same data is analysed (and reported) using the latter.

9.3.3. Different measures of the predictor variables

Another potential issue is that the (predictor) variables need to be measured similarly in each analysis. In the simplest case, this means that some effect sizes need to be rescaled if the reported effects are based on different quantities of the same units. This is evident from how discrete choice odd-ratio effects represent the effect of a one-unit change in the predictor variable. As such, if the same predictor variable is measured in 10s and 100s of units respectively, then the estimated effect in the second analyses will appear (relatively) 10 times larger. This, however, can be fixed by rescaling the estimated effect in either of the studies. A similar issues arises when analyses use opposite units to other analyses (e.g. proximity compared to distance). These are possible if the difference in the units is a linear transformation. However, elsewhere such as when distance and log(distance) are used, the effects are (generally) not combinable because taking the logarithm is a non-linear transformation. Here, and particularly given the difference in results (e.g. in Johnson and Summers, 2015), these must be treated as distinct variables.

A similar issue is that some variables are calculated using different methods. These differences can be relatively inconsequential. For example, the residential mobility/churn variable is calculated in a number of ways. These include the annual percentage of residents who moved into or out of a neighbourhood (e.g. Bernasco and Nieuwbeerta, 2005), the annual percentage who moved into, out of or within a neighbourhood (e.g. Baudains, Braithwaite and Johnson, 2013) and using the index of qualitative variation (Frith, Johnson and Fry, 2017). Another example is where some variables such as the
presence of schools are operationalised as the number of schools in an area (e.g. Lammers et al., 2015) or as a binary variable to indicate if a school is present or not (e.g. Bernasco and Block, 2009). In both examples (and others), the effects should be reasonably similar regardless of operationalization and very similar results were found when the different versions of the same variables were tested in analyses of the available raw data (e.g. Frith, Johnson and Fry, 2017). These minor differences can therefore be ignored with little or no consequence.

However, for two sets of variables the differences in operationalization was much greater. The first was for the measurement of the level of affluence in an area. For this, authors have used measures including the simple average household value (e.g. Bernasco and Nieuwbeerta, 2005), factor variables (e.g. loaded on household value, annual income and the percentage of properties that are owner-occupied in Bernasco 2006), country-specific composite indices of income deprivation (Summers, 2012), and socio-economic disadvantage (Clare et al., 2009).

The second is for the measurement of social disorganisation and collective efficacy for which authors have used individual variables traditionally associated with social cohesion such as residential mobility and ethnic heterogeneity (e.g. Bernasco and Nieuwbeerta, 2005), factor-load combinations of those variables (e.g. Bernasco, 2006) or variables from surveys specifically designed to elicit perceived levels of social disorganisation (e.g. Bernasco and Block, 2009; see also Sampson, Raudenbush and Earls, 1997). For these, the variables are meta-analysed if their specific measurement is repeated in multiple analyses (in the case of household value, residential mobility and ethnic heterogeneity) and the remaining idiosyncratic (composite) variables for each are all meta-analysed together though this potential source of inconsistency is considered when interpreting the results.
9.3.4. Non-independent datasets

The next issue regards the non-independence of the data where the same or similar datasets are used in multiple analyses. As described and illustrated earlier, this includes cases where multiple analyses use the exact same datasets. It also includes analyses that employ overlapping datasets. For example, where some analyses use all offences in a dataset whereas others use a subset. It also includes cases where some analyses use data for one time-period (in a study area) and other analyses use data from an overlapping time-period (for the same study area) such that both analyses (likely) contain some of the same offences. One consequence of this is that the effects from these analyses will correlate. As such, because standard meta-analyses assume the effects are independent (across analyses) and each analysis provides new information about the effects, the meta-analysis will overestimate the amount of information regarding each effect. In statistical terms, the standard errors for each effect will be incorrect and too small (Becker, 2000). Ignoring this correlation will also mean that the results from the meta-analysis will be biased towards those datasets that have been analysed on more occasions.

In general, there are three general strategies for dealing with sample dependence (Becker, 2000; Van den Noortgate et al., 2013). Firstly, the dependence can be ignored. In the case of this literature, where some datasets are analysed on many (e.g. 12) or few (e.g. 1) occasions this is in principle not appropriate. The second strategy is to avoid the dependence, for example, by including only one analysis per dataset. This will however exclude a large number of analyses and their heterogeneity (e.g. the inclusion or not of a variable) which can provide useful data on the effects of those variables (including the effect of the inclusion of a variable on the effects of the other variables). This strategy would also be poor for dealing with the dependence from analyses of partially overlapping
samples as they would need to be removed (to avoid any kind of dependence) or potentially included (and so only some partial dependence is removed).

The third and more complex (but more accurate) strategy is therefore followed (in the meta-analyses and meta-regressions) whereby the dependence (and partial dependence) is modelled using a multilevel or hierarchical parameterisation of the meta-analysis (and meta-regression) model. In a multilevel meta-analysis, this dependency is accounted for by grouping the analyses into nests based on the dataset used. The correlation within and between those nests are then specified based on the overlap of the data in the analyses using the following formula:

\[ \text{Corr}_{ab} = \sqrt{k_{a,b}k_{b,a}} \]

Where \( a \) and \( b \) are a pair of datasets, \( k_{a,b} \) is the proportion of the sample in dataset \( a \) that is in dataset \( b \) and \( k_{b,a} \) is the proportion of the sample in dataset \( b \) in dataset \( a \) (Steel and McLaren, 2009; see also Tam, 1984; Laniel, 1988). To give two examples, first consider the two analyses in Bernasco and Block (2009) that used the same dataset. Here, \( k_{a,b} \) and \( k_{b,a} \) will both equal one and so the effects from those analyses are calculated to perfectly correlate (a correlation coefficient of 1). This perfect correlation obviously relies on all else (between those two analyses) being equal whereas the analyses should differ in some way (e.g. the inclusion or not of a variable). In fact, this knowledge that the estimates should be equal will help the meta-analyst determine the cause and effect (including its magnitude) of the variation in the estimates. In a more complicated example

---

33 In the original formula, \( k_{a,b}k_{b,a} \) is also multiplied by \( r_{ab} \) which is the individual-level correlation between values but as these data involve the same choices, \( r_{ab} \) will equal 1 and so is ignored in the formula presented here.

34 If the exact same analyses are reported on multiple occasions, the extra duplicate analyses add no information to the meta-analysis but will cause the nesting structure to be more complicated (e.g., an extra nesting level may need to be added). As such, these duplicate analyses are ignored.
consider the datasets (of burglaries in The Hague) used in some of the analyses in Bernasco and Nieuwbeerta (2005) and Bernasco (2006). In the former analysis the dataset, now referred to as \( a \), contains 548 burglaries for the time-period 1996 to 2001. In the latter dataset, now referred to as \( b \), there are 809 burglaries from the (overlapping) period of 1996 to 2004. Due to a lack of information of the exact observations (offences) included in either dataset, it is assumed for simplicity that the offences are equally distributed throughout each respective time-period. As such, all 548 offences from \( a \) are in \( b \) (\( k_{a,b} = 1 \)) and 539 (\( \frac{809}{9} \times 6 \)) of the offences in \( b \) are in \( a \) (\( k_{b,a} = 0.67 \)) and so the correlation is estimated at 0.82.

9.3.5. Other analysis differences

As described earlier, there are also other differences between analyses that may need to be accounted for. The most notable difference and most likely cause of heterogeneity between analyses is the types of offences being analysed. For example, and although no hypotheses are explicitly made, because discrete choice analyses assume the offence (location) choices involve some element of rationality (i.e. they chose locations where they expect to derive the greatest utility) it is possible these methods are more suited to offences likely to be deliberate and ostentatiously purposive such as acquisitive offences (e.g. Cornish and Clarke, 1986). In contrast, analyses of non-acquisitive offences (e.g. violence), which are more likely to involve an emotional or expressive element, may detect smaller influences (or no influence at all) of the various variables. As such, although the meta-analyses and meta-regressions will be used to first synthesise all analyses (regardless

---

35 The mix (or not in the case of \( b \)) of single-offender and group-offender burglaries is ignored for simplicity.

36 Minor to moderate differences will have little overall effect. For example, even if the numbers of crimes in each dataset that are in the other dataset are rounded down to the nearest 100 (to 500 for both) the correlation is estimated as 0.75.
of offence type) together, if substantial heterogeneity in the effects is detected (see 7.3.7) and plausible and definable sub-types of offences are present (e.g. all offences can be split into those that are acquisitive and non-acquisitive) then separate meta-analyses and meta-regressions will be conducted on each sub-type.

There may also other differences between the analyses including the location of the study area (in terms of the country\textsuperscript{37}), the size of the spatial units used in the analyses, and whether the analysis was published (in terms of in a peer-reviewed journal or book) which may affect the estimates in each analysis. To give an example for each; firstly, given the dependency on cars in the US and the opportunity cars offer to travel further distances in shorter times (e.g. see the analysis of travel surveys in Giuliano and Dargay, 2006), distance (to possible offence locations) may have a less prohibitive effect in the US than other countries. Second, the size of the spatial unit may affect the results. Here, if offenders follow a hierarchical spatially structured location decision process, they may use different criteria or the influence of those criteria may vary at different scales. As such, analyses at different scales may find different results for the same variables. Lastly, whether an analysis is published or not may relate to the fact that published analyses generally have larger effect sizes as they are more likely to be submitted and published (e.g. Franco, Malhotra and Simonovits, 2014). As such, although these differences are ignored in the meta-analysis, in the meta-regression they are accounted for (and hypothesis tested for their effect) using moderator variables. Note that unlike the other moderator variables (including below) which are included as binary indicators, the size of

\textsuperscript{37} The city or general area (within the country) could also be included but is not due to the number of moderator variables that would need to be included.
the spatial unit is included as a continuous covariate based on its size in km² and 1,000s of residents.\textsuperscript{38}

9.3.6. Other issues

The first other issue regards the fact that some variables are only included in a few analyses or only in analyses of overlapping datasets. As such, little information is available regarding the true effects of those variables. For this reason, variables that have their effects reported in fewer than five analyses and for only one study area are omitted. In doing so, the meta-analyses and meta-regressions only focus on the (18) most commonly used variables. Also, note that for sets of interaction variables that would be omitted, which would include the effect of distance to the city centre for juvenile and adult offenders, these are combined using a fixed-effect model to derive an estimate of the overall effect (across juvenile and adult offenders).\textsuperscript{39} This is then included if that (overall) variable meets the criteria above.

Next, some analyses (and the results from those analyses) are excluded if the data used in the analysis is atypical and non-random. For example, all 19 analyses in Bernasco et al, 2017 are for offences from the different periods of the day and different days of the week. These are omitted here because the estimated effects from those analyses will not be

\textsuperscript{38} Note that including the size of the spatial unit in terms of binary variables for small, medium and large (see also earlier) did not substantially affect the results and so is not discussed in the results.

\textsuperscript{39} The efficacy of this process can be demonstrated using data from Townsley et al. (2015) and meta-analysing the separate CL estimates of the effect of distance (from the offender’s home) for adult and juvenile offenders and comparing it to the CL estimated effect of distance in the equivalent CL model without the distinction between offenders. Here, the meta-analysis combined estimates were 0.58 (se=0.05), 0.52 (se=0.04) and 0.82 (se=0.02) for The Hague, Birmingham and Brisbane study areas compared to the CL model-derived estimates of 0.60 (se=0.05), 0.53 (se=0.03) and 0.83 (se=0.02).
representative of the general distribution of effects (i.e. in other study areas rather than in those specific subsets of offences).

A third issue solely related to the meta-regressions, regards the relatively large number of moderators (up to 25) that can be included in each meta-regression. For one, including a large number of moderator variables increases the risk of multicollinearity and can cause unstable estimates of the effects of those variables. As such, for each meta-regression, multicollinearity is investigated using the Pearson correlation coefficient (for continuous variables) and Phi coefficient (for binary variables) and moderators are excluded should they exceed ±0.70 (e.g. Dormann et al., 2013). Including too many moderator variables than can be justified by the quantity of data can also induce over-fitting. This is where the meta-regression model describes the random error in the data and so can yield misleading results. Here, excessive moderators, defined as where there are more than five covariates per estimate to be combined (Harbord and Higgins, 2008), are removed by repeatedly re-estimating the meta-regression and omitting the covariate with the smallest p-value until a satisfactory number of covariates per estimate remain\textsuperscript{40}.

9.3.7. Model specifications

The meta-analyses and meta-regressions are estimated using the ‘metafor’ package (Viechtbauer, 2010) and the recommended restricted maximum likelihood method (e.g. Thompson and Sharp, 1999; Viechtbauer, 2005) in the statistical software environment R (R Core Team, 2016).

\textsuperscript{40} Other options, such as a criterion of 10 estimates per covariate and including all covariates, were also tested though the results from these (regarding the effect of the covariates) are relatively similar and so are not discussed further.
9.4. Meta-analysis results

Shown in Figure 9.4 are the results, where applicable (see earlier), from the meta-analyses of the estimates for each of the 18 variables from all of the discrete choice analyses (regardless of the offence type). Different estimates are shown for acquisitive and non-acquisitive offences (regardless of the offence sub-type), and for the separate analyses of each individual offence (sub-)type. Plotted in Figure 9.4 are the estimated average effects for each variable (the black square) and its 95% confidence intervals (the solid line whiskers) and prediction intervals (the dashed whiskers) (see Chapter 4 for more information for these metrics). These are also shown in the summary data in Figure 9.4 along with the number of estimates contributing to each meta-analysis (I), the significance of the hypothesis test that the mean estimate significantly differs from no effect is also shown, along with the significance of the \( Q \)-test of heterogeneity and the \( I^2 \) measure of heterogeneity (see Chapter 4 for more information for these metrics).

The results show that, on average, many of the variables tested influence the offence location choices for many of the different types of offenders. In the case of distance, all estimates were statistically significant (\( p < 0.05 \)), with there being a negative association between distance and location choice. With the exception of non-acquisitive offences (\( p > 0.05 \)), this was true for all (analysed) offence types and sub-types. Wald tests to compare the mean effect of distance for juvenile and adult offenders found a significant difference, where juvenile offenders are, on average, more restricted by distance, when including all types of offences (\( b = -0.17, SE = 0.08, p < 0.05 \)). However, though a similar difference was found for acquisitive offenders (\( b = -0.17, SE = 0.09 \)), the difference was (marginally) non-significant (\( p = 0.06 \)).
Figure 9.4: Forest plot of the findings from meta-analysis results
The other factors found to be, on average, significant positive predictors of an area being selected for an offence were: being the offender’s home area (for all offenders and for acquisitive offenders); the number of targets (but only for all/acquisitive and robbery offenders); estimated levels of social disorganisation (for all and acquisitive offenders), ethnic heterogeneity (for all and acquisitive offenders) and residential mobility (but only for all and acquisitive offenders); a river between it and the offenders’ home (but only for acquisitive offenders), a main street being present (for all and acquisitive offenders); and the presence of trains/bus stations (for all and acquisitive offenders) and them being connected to train/bus stations in the offender’s home area (for all/acquisitive offenders) (all $p < 0.05$). A main road located between the area and the offender’s home (for all/acquisitive offenders) was the only variable (along with the distance variables, see above) found to be a significant negative predictor ($p < 0.05$). The remaining variables (distance to the city centre, affluence (composite or household value) and presence of schools) were not found to be, on average, significant predictors of offence locations.

While the estimated mean effects of each variable can be compared across offence types and sub-types (e.g. burglary offenders to robbery offenders), due to relatively few studies analysing either non-burglary or non-acquisitive offences there are few opportunities. The results of the possible comparisons are shown in Table 9.4 though they must be interpreted cautiously due to the relatively low number of studies used to derive each estimate. Here, there is no evidence in these comparisons (of the effects of distance, distance to the city centre, the number of targets and residential mobility) that burglary offenders (statistically significantly) differ to the general group of other types of non-acquisitive offenders (three variables) or to robbery offenders (one variable: number of

---

41 If using the same criteria as applied for the meta-analysis or meta-regression - that there must be at-least five estimates of the effect of a variable and that these must be from at-least one separate study area.
targets); or that acquisitive offenders differ to non-acquisitive offenders (one variable: distance).

Whilst these comparisons are for the mean effect of each variable, the results in Figure 9.4 also show that there is often a relatively large amount of heterogeneity in the effects (sizes) for each variable that are combined in the meta-analyses. More specifically, in 30 of the 46 meta-analyses the Cochran’s $\chi^2 Q$-test, which assesses whether the observed differences in the effect sizes are compatible with chance, is significant at $\alpha = 0.05$. Additionally, the $I^2$ index which measures the percentage of variation that is due to heterogeneity is high in 36 of the 46 meta-analyses and medium-to-high in a further five. The only meta-analyses for which there was no evidence of large amounts of heterogeneity were for the number of targets for robbery offenders, the effect of ethnic heterogeneity for all and acquisitive offenders, and the effect of a main street for all and acquisitive offenders. For the other variables though, these measures indicate that the effects (on offence location choices) vary across the analyses. This is illustrated in Figure 9.4 by the tau-derived prediction intervals (the grey dashed lines) which show the range expected to contain the effect size of 95% of each variable in the (potential) future discrete choice analyses. For example, and based on the current analyses (see below), the odds-

<table>
<thead>
<tr>
<th>Variable compared</th>
<th>Types of offenders compared</th>
<th>N</th>
<th>Effect estimate (b, SE)</th>
<th>Wald test comparison result (b, SE, p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (km)</td>
<td>Acquisitive offenders</td>
<td>23</td>
<td>-0.34, 0.04</td>
<td>0.12, 0.14, p&gt;0.05</td>
</tr>
<tr>
<td></td>
<td>Non-acquisitive offenders</td>
<td>5</td>
<td>-0.22, 0.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Burglary offenders</td>
<td>12</td>
<td>-0.33, 0.07</td>
<td>-0.04, 0.09, p&gt;0.05</td>
</tr>
<tr>
<td></td>
<td>Non-burglary acquisitive offenders</td>
<td>9</td>
<td>-0.37, 0.07</td>
<td></td>
</tr>
<tr>
<td>Distance to the city centre (km)</td>
<td>Burglary offenders</td>
<td>8</td>
<td>0.00, 0.04</td>
<td>0.06, 0.06, p&gt;0.05</td>
</tr>
<tr>
<td></td>
<td>Non-burglary acquisitive offenders</td>
<td>7</td>
<td>0.05, 0.05</td>
<td></td>
</tr>
<tr>
<td>Number of targets (100)</td>
<td>Burglary offenders</td>
<td>11</td>
<td>0.58, 0.42</td>
<td>-0.54, 0.42, p&gt;0.05</td>
</tr>
<tr>
<td></td>
<td>Robbery offenders</td>
<td>8</td>
<td>0.04, 0.01</td>
<td></td>
</tr>
<tr>
<td>Residential mobility (10%)</td>
<td>Burglary offenders</td>
<td>8</td>
<td>0.08, 0.06</td>
<td>0.12, 0.08, p&gt;0.05</td>
</tr>
<tr>
<td></td>
<td>Non-burglary acquisitive offenders</td>
<td>5</td>
<td>0.20, 0.05</td>
<td></td>
</tr>
</tbody>
</table>
ratio multiplicative effect of an increase in the number of targets in an area on the likelihood of a burglar offending in that area in 95% of the future analyses (of burglary location choices) is estimated to be between the extremes of 0.06 and 51.94. In contrast, the multiplicative effect of an increase in distance of 1km from the burglar’s home in future analyses is estimated to be between 0.59 and 0.82.

While large (and in five cases, less than large) amounts of heterogeneity were found for the estimates of the effects of each variable (and so wider or smaller ranges are predicted for any future analyses), two important caveats should be considered. Firstly, although the estimated mean effects, etc. concern all future analyses regardless of its setting, if the estimates for a variable are predominantly (or only) from analyses of one type of study (e.g. one study area, spatial unit size, offence sub-type), then the resulting estimate may only or largely reflect future analyses of that study type. The key examples of this are for ethnic heterogeneity, for which 88% of all estimates were from the UK. In addition, where 80% of the train/bus station estimates and 83% of the ‘is the offender’s home neighbourhood’ estimates are from the Netherlands.

Secondly, and in a similar way, because the estimated effects likely vary to some degree depending on the study type (e.g. the study area, spatial unit size, the other variables included in the analyses), then the estimates will likely appear more or less heterogeneous depending on the presence and variability of analysis-level differences. Possible effects from these analysis-level differences can be estimated and accounted for using a meta-regression (see below) – although for many of the combinations of variables and offender types there are insufficient estimates for a meta-regression to be conducted.
9.5. Meta-regression results

Figure 9.5 shows the results for the meta-regressions. As before, the estimates are shown for all discrete choice analyses (regardless of the offence type), for the separate analyses of acquisitive and non-acquisitive offences (regardless of the offence sub-type), and for the separate analyses of each individual offence (sub-)type. Plotted in the figure are the results for the meta-regression (black) and, for the purposes of comparison, the equivalent findings from the above meta-analysis (grey; see also Figure 9.4). It shows their estimated average effect (the square), 95% confidence (the solid line whiskers) intervals and prediction intervals (the dashed whiskers). These are shown alongside the number of estimates used in each analysis (N), the significance of the test that the mean estimate significantly differs from no effect, and the significance of the $Q$-test of heterogeneity and $I^2$ measure of heterogeneity. For the meta-regressions, the number of moderators included in the analysis and the number found to significantly affect the mean effect of each variable, the significance of the $Q$-test that the moderators influence the effect and the $R^2$ measure of the percentage of heterogeneity between estimates that is explained by the moderators are also shown.

Similar to the results from the meta-analysis, these results show that after accounting for the effect of (up to 25) analysis-level differences, a large number of variables still, on average, appear to influence the offence location choices of many of the different types of offenders. In terms of the distance (from the offender’s home) variables, they all (that could be analysed) had a significant negative effect on the likelihood of a location being selected (all $p < 0.05$). In line with the previous results, when considering all types of offenders, distance had a greater impact for juvenile offence location decisions ($b = -0.24, SE = 0.06, p < 0.01$) than it did for adults. Moreover, unlike the results of the meta-analysis, after accounting for analysis-level differences, the same was true for
Figure 9.5: Forest plot of the findings from meta-regression (black) and meta-analysis (grey)
## Systematic review and meta-analysis of offence location choice research

### Odds-ratio

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>0.1</th>
<th>0.25</th>
<th>0.3</th>
<th>0.25</th>
<th>1</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>25</th>
</tr>
</thead>
</table>

### Number of targets (136)

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.14</td>
<td>(0.07, 0.31)</td>
<td>100%</td>
<td>0.01</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.97</td>
<td>(0.94, 0.99)</td>
<td>100%</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Social disorganisation (composite)

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.50</td>
<td>(0.41, 0.60)</td>
<td>76%</td>
<td>0.04</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.52</td>
<td>(0.43, 0.61)</td>
<td>76%</td>
<td>0.04</td>
</tr>
</tbody>
</table>

### Ethnic heterogeneity (25%)

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.06</td>
<td>(0.00, 0.35)</td>
<td>0%</td>
<td>0.05</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.07</td>
<td>(0.03, 0.24)</td>
<td>0%</td>
<td>0.05</td>
</tr>
</tbody>
</table>

### Residential mobility (50%)

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1.00</td>
<td>(1.00, 1.00)</td>
<td>100%</td>
<td>0.01</td>
</tr>
<tr>
<td>Acquittal</td>
<td>1.00</td>
<td>(1.00, 1.00)</td>
<td>100%</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### River barrier between target area

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.17</td>
<td>(0.09, 0.32)</td>
<td>68%</td>
<td>0.02</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.16</td>
<td>(0.08, 0.32)</td>
<td>68%</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Main street in target area

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.55</td>
<td>(0.41, 0.73)</td>
<td>82%</td>
<td>0.01</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.55</td>
<td>(0.41, 0.73)</td>
<td>82%</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Presence of school in target area

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.41</td>
<td>(0.34, 0.50)</td>
<td>98%</td>
<td>0.02</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.41</td>
<td>(0.34, 0.50)</td>
<td>98%</td>
<td>0.02</td>
</tr>
</tbody>
</table>

### Train/bus connect target area

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.42</td>
<td>(0.34, 0.50)</td>
<td>84%</td>
<td>0.01</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.42</td>
<td>(0.34, 0.50)</td>
<td>84%</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Train/pass station in target area

<table>
<thead>
<tr>
<th>Group</th>
<th>Odds Ratio</th>
<th>Confidence Interval</th>
<th>Q-test</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.34</td>
<td>(0.28, 0.41)</td>
<td>89%</td>
<td>0.1</td>
</tr>
<tr>
<td>Acquittal</td>
<td>0.34</td>
<td>(0.28, 0.41)</td>
<td>89%</td>
<td>0.1</td>
</tr>
</tbody>
</table>

---

- 236 -
juvenile acquisitive offenders compared to their adult counterparts \((b = -0.23, SE = 0.09, p < 0.05)\). There was no difference for burglary offenders \((b = -0.14, SE = 0.20, p > 0.05)\). The other factors found to be, on average, significant predictors of offence location choices are: a location encapsulating the offender’s home neighbourhood, distance to the city centre (but only for all offenders), affluence (the composite variable) (but not for burglary offenders), social disorganisation, residential mobility, and the presence of a school in the target area (but only for all offenders).

These meta-regression results are generally consistent with those of the meta-analysis. This is despite the fact the meta-regression attempts to account for study-level differences between analyses such as the inclusion or not of variables. However, there were some differences. The key between the two sets of results are for the negative effects of distance (km) for which the estimated effect was larger for the meta-regression (all \(p < 0.01\)). In contrast, the effect of logged distance (km) appears significantly weaker for the meta-regression \((b = -0.76, SE = 0.18, p < 0.01)\). Similarly, the meta-regression identified distance to the city centre and affluence (the composite variable) to be significant negative predictors of offence locations (for all offenders for the former, and for all and acquisitive offenders for the latter) whereas the meta-analysis suggested they had no significant effect. The meta-regression also suggested that residential mobility had a significantly greater positive effect on offence location choices than would be suggested by the findings of the meta-analysis for all offenders \((b = 0.87, SE = 0.23, p < 0.01)\) and acquisitive offenders \((b = 0.99, SE = 0.24, p < 0.01)\).

In terms of comparing the average effects (from the meta-regressions) of the variables across offender types, Table 9.5 shows that consistent with the comparisons from the meta-analyses (see also Table 9.4), burglars do not significantly differ from the other types of acquisitive offenders in terms of preferring offence locations closer from the city centre.
or having greater residential mobility (both $p > 0.05$). Unlike the meta-analysis, the meta-regression however found that burglars have a greater preference for closer offence locations compared to the other types of acquisitive offenders ($b = 0.28, SE = 0.08, p < 0.01$).

Table 9.5: Results from Wald tests comparing effect estimates from the meta-regression

<table>
<thead>
<tr>
<th>Variable compared</th>
<th>Types of offenders compared</th>
<th>N</th>
<th>Effect estimate ($b, SE$)</th>
<th>Wald test comparison result ($b, SE, p$-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (km)</td>
<td>Burglary offenders</td>
<td>18</td>
<td>-0.64, 0.05</td>
<td>0.28, 0.08, $p &lt; 0.01$</td>
</tr>
<tr>
<td></td>
<td>Non-burglary acquisitive offenders</td>
<td>13</td>
<td>-0.36, 0.06</td>
<td></td>
</tr>
<tr>
<td>Distance to the city centre (km)</td>
<td>Burglary offenders</td>
<td>18</td>
<td>-0.04, 0.04</td>
<td>0.09, 0.05, $p &gt; 0.05$</td>
</tr>
<tr>
<td></td>
<td>Non-burglary acquisitive offenders</td>
<td>14</td>
<td>0.05, 0.03</td>
<td></td>
</tr>
<tr>
<td>Residential mobility (10%)</td>
<td>Burglary offenders</td>
<td>17</td>
<td>0.17, 0.05</td>
<td>0.04, 0.06, $p &gt; 0.05$</td>
</tr>
<tr>
<td></td>
<td>Non-burglary acquisitive offenders</td>
<td>10</td>
<td>0.21, 0.03</td>
<td></td>
</tr>
</tbody>
</table>

While including the moderators in the meta-regression accounts for the analysis-level differences in the estimates of the effects of each variable, the effect of each analysis-level difference on those estimates can also be investigated. These are summarised in Figure 9.6. Dark grey cells indicate that the analysis-level difference (moderator) was excluded due to insufficient studies using that moderating variable, to avoid multicollinearity or over-fitting. White cells indicate that the moderator had no significant effect in the final model, and green (red) cells indicates the moderator had a significant positive (negative) effect on the mean effect of each variable. From the results shown, certain patterns regarding the effects of the moderators are noteworthy, for example, either because they have a (relatively) consistent effect across offence types or because they have a substantial effect.

The first notable result is the (relatively) consistent positive effect of an analysis being conducted on offenders in Australia on the effects of distance (from the offender’s home). That is, while distance generally constrains offence location choices in Australia (and
**Figure 9.6: Euler grid diagram showing the meta-regression results including the average effects of each variable and the effects of each moderator variable on the effects of each variable**

<table>
<thead>
<tr>
<th>Offence Type</th>
<th>Focal Variable</th>
<th>Average effect (meta-regression)</th>
<th>Effect of each moderator variable</th>
<th>Effect of each moderator variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>Distance (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Distance (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Distance (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Distance (log km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Distance (log km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Distance (log km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Distance (juveniles) (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Distance (juveniles) (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Distance (juveniles) (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Distance (adults) (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Distance (adults) (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Distance (adults) (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Distance to city centre (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Distance to city centre (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offence Type</td>
<td>Focal Variable</td>
<td>Average effect (meta-regression)</td>
<td>Is in Netherlands</td>
<td>Is in US</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
<td>----------------------------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Burglary</td>
<td>Distance to city centre (km)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Is target area their own area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Is target area their own area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Affluence (composite)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Affluence (composite)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Affluence (composite)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Affluence (house value) (£10,000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Affluence (house value) (£10,000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Affluence (house value) (£10,000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Number of targets (100)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Number of targets (100)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Social disorganisation (composite)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Social disorganisation (composite)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Ethnic heterogeneity (10%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Ethnic heterogeneity (10%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Ethnic heterogeneity (10%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offence Type</td>
<td>Focal Variable</td>
<td>Average effect (meta-regression)</td>
<td>Is in Netherlands</td>
<td>Is in US</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------</td>
<td>----------------------------------</td>
<td>-------------------</td>
<td>---------</td>
</tr>
<tr>
<td>All</td>
<td>Residential mobility (10%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Residential mobility (10%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Burglary</td>
<td>Residential mobility (10%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>River barrier between target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>River barrier between target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Road barrier between target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Road barrier between target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Main street in target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Main street in target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Presence of school in target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Presence of school in target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Train/bus connect target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Train/bus connect target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>Train/bus station in target area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acquisitive</td>
<td>Train/bus station in target area</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Column 3 indicates the average effect of each variable in the respective offence location choice analysis; the remaining columns (on its right) indicate the effect of each variable or analysis-level difference on the average effect of each variable. Black indicates the moderator variable was omitted due to insufficient studies, multicollinearity or to prevent over-fitting. White indicates no significant effect, green indicates a significant positive effect, red indicates a significant negative effect.
elsewhere), offenders in Australia are less inhibited by distance and so are more willing to travel further distances. For example, in terms of distance (km), the results suggest that for an average offender for every kilometre a location is further away, the odds of it being selected decreases by 0.40 ($b = -0.91, p < 0.01$) whereas for offenders in Australia it only decreases by 0.52 ($b = -0.66, p < 0.01$). In contrast juvenile offenders in the UK were more constrained by distance than their non-UK counterparts.

There are also three other consistent moderators (specifically) related to the distance variables. First is that excluding the variable which measures the distance to the city centre reduces the estimated effect of the distance between the offender’s home and offence location. In this case, the odds-ratio for an average offender reduces from 0.40 ($b = -0.91, p < 0.01$) to 0.55 ($b = -0.60, p < 0.01$) when the distance to the city centre is omitted. Second, the omission of estimates of social disorganisation tends to exaggerate the effect of distance (log km). For example, for an average offender, the odds-ratio for distance (offender home to offence location) increases from 0.30 ($b = -1.21, p < 0.01$) when estimates of social disorganisation are included in the analysis to 0.22 ($b = -1.53, p < 0.01$) when they are not. Third, omitting the presence of main streets in the analysis appears to exaggerate the negative effects of distance for adult offenders (in general) and adult acquisitive offenders.

---

42 Here, and elsewhere, the modelled estimates are based on a discrete choice analysis of offenders (regardless of offence type and study area location) that, where applicable and unless otherwise stated, includes all variables, is published and is calculated using MI, using the mean sized spatial units (an average area of 2.7km² and with 2,775 residents).
Interestingly, the results also indicate that its estimated effect for all offenders and acquisitive offenders also depends on the size of the spatial units being used. That is, while higher levels of residential mobility generally increase the likelihood a spatial unit will be selected for an offence, when the spatial units being used are larger this effect tends to diminish and potentially become negative. That said, the latter finding (become negative) might result from assuming a linear relationship between size and effect. According to the results, unpublished analyses of all offenders, acquisitive offenders and burglary offenders are likely to find smaller effects of residential mobility. One other common effect concerns the statistical model used. If CL rather than ML is used, the effect of distance (km) appears to be understated. Using the CL model also has a consistent positive effect on the estimated effects of distance to the city centre in the analyses for all, acquisitive and burglary offenders.

9.6. Discussion

The aim of this chapter was to formally collate and synthesise the findings of current research on offender location choices that employs the discrete choice approach. To do this, a systematic review of the literature was conducted. The results were then pooled using a meta-analysis. However, as there were analysis-level differences, a meta-regression was also conducted which can be accounted for these using moderator variables. That being said, due to the lack of data the meta-regression could only be computed for some variables. Both approaches are therefore used together.

Previous studies of offence location choice analyse data for different samples of offenders separately. As such, they only provide estimates of the influence of variables for the particular types of offenders considered and the areas in which they offend. In the current study, data were synthesised across many studies and types of offenders allowing hypotheses to be tested about the external validity of finding. The results suggest many
of the variables consistently affect (or not) offence location choices across study areas and offender types. For example, in line with the rational choice perspective (see Chapter 2), the results indicate that all offenders, regardless of type or location, tend to select locations that are nearer to their home location than those that are further apart. This finding was also robust regardless of how distance was measured (i.e. in terms of kilometres, the logarithm of kilometres) or if distance (in kilometres) was examined for juvenile and adult offenders separately. Other variables which tended to have a consistent effect were the distance to the city centre, if the location is where the offender lives, some measures of affluence and social disorganisation, residential mobility and the presence of schools.

That being said, one important finding from these analyses is that there is heterogeneity in the average effects, or the offenders’ preferences, of the variables (between analyses). That is, whilst Townsley et al. (2015b) and Frith et al. (2017) show there is preference heterogeneity between offenders within the same sample, the results in this analysis show that there is also heterogeneity in the average preferences between samples. This heterogeneity exists even after accounting for the differences that would be expected to exist due to the fact that the data concern different types of offences and sub-types of offenders. That said, for example, as shown in Table 7 (see also $I^2$ values in Figures 4 and 5), for burglars, there was less heterogeneity - as indicated by there being are slightly fewer statistically significant Q-tests (and generally slightly smaller $I^2$ values) - than for the analyses of all offenders. This suggest that, to some degree, different types of offenders behave somewhat similarly in terms of the factors that influence their offence location choices – though this requires further research to explore and elaborate (but is beyond the scope of this thesis).

The results also indicated that context matters in terms of offenders in one country, on average, systemically differing in their offending preferences to those in another
Two examples illustrate this. First, offenders in Australia appeared to be less inhibited by distance than offenders elsewhere. Second, juvenile burglary offenders in the UK appeared to be far more constrained by distance than their non-UK counterparts. A possible explanation for the former is that there is a much lower density of targets (people or households per km²) in Australia than in most other (included) countries. For example, in 2016 in the UK there were on average 269.6 residents (ONS, 2017) and 111.2 households (ONS, 2016) per km² whereas in Australia the equivalent figures are 3.0 residents and 1.3 households per km² (Australian Bureau of Statistics, 2018). As such, offenders in Australia will likely, on average, have to travel further to find suitable targets. For the latter this could be explained as driving, or at-least holding a driving license and so having the (official) ability to drive, tends to be more prevalent in other countries than the UK. For example, in 2015 in the UK ~0% of 16 year olds and 27% of 18 year olds hold a driving license (DVLA, 2015; ONS, 2015) whereas in the US it is around 25% of 16 year olds and 61% of 18 years old (U.S. Department of Transportation, 2016). Therefore, and for the same reasons, juvenile offenders are expected to be more constrained by distance than adult offenders (e.g. see Bernasco and Nieuwbeerta, 2005).

These potential explanations require testing alongside other possibilities.

The size of the spatial units (in terms of number of residents and square area) used in the analyses were also found to impact on the results for some variables. Notably, when spatial units are larger in size, the influence of residential mobility (for all offenders and acquisitive offenders) on offence location choices was generally weaker (and eventually became negative). In addition, for larger spatial units, the effects of schools in an area (for all offenders) was larger. Though these scaling effects likely exist to some degree, in that here size was incorporated as continuous linear effect whereas it likely stops at some point (i.e. schools in an infinitely sized spatial unit would not infinitely effect offence location
choices), this finding follows from results in other analyses (e.g. Oberwittler and Wikström, 2009) and can be explained in two ways. Firstly, theoretically, whereby even if some variables have some effect (on offence location choices), unless they are incorporated at a spatial scale which reflects how offenders sense them or affects them, their effect can be mis-estimated. For example, whilst the turnover of residents in the near vicinity of a target may make that target more attractive to offenders, turnover (or lack thereof) in areas that are more distant will likely have less or no effect. Secondly, methodologically, whereby larger spatial units are more likely to contain heterogeneous environmental conditions and so the measures of those conditions at those scales is likely to be misrepresentative. As such, even if again those conditions affect offence location choices, the statistical models (discrete choice models or others) may be unable to detect their correct relationship. Even if either explanation (or any other) is true in the case of these analyses, these findings nonetheless highlight the importance and need for further research into the applicability of spatial units, multi-level models and/or spillover variables.

Lastly, the results highlight the importance of model specification and in particular, under-specification. That is, whilst certain independent variables should be included in the analyses because they appear or are expected to (directly) effect offence location choices, there are also other variables that have indirect effects and so should be included to adjust for their effect and prevent the misestimation of coefficients. For example, while only distance and residential mobility were found to be significant predictors of burglary offence locations in the meta-regressions, the results for these and other variables are estimated to be biased, to some degree, if the distance to the city centre, affluence and the offender's previous offence locations are not also included in the analyses.
Burglary Offence Location

Choice Analyses
Role of guardianship and familiarity in burglary location choices

This thesis has so far discussed the likely influences of burglar location choices and established that movement, either in terms of the offender’s or passers-by’s, should play a large role in where burglaries occur. In Chapter 5, several novel graph theory metrics were created to estimate different aspects of this movement. In this chapter, using the discrete choice approach introduced and reviewed in Chapters 6-9, these metrics are used to test hypotheses motivated by theory. The main contribution of this chapter is to determine whether the novel metrics warrant further investigation and if they should be considered in future analyses, in particular, in the analysis of burglar location preference variation, which is presented in Chapter 11.

10.1. Introduction

In this thesis, I have so-far discussed the likely influences of (burglary) crime location choices and established that movement, either in terms of offenders or passers-by, should play a role in where (burglary) crimes occur. Most criminological research has however failed to (adequately) incorporate estimates of movement in their analyses (Chapter 3).
This is despite the fact that there exist suitable measures, graph theoretical measures and specifically the betweenness measure, which correlate with overall (passers-by) movement (Chapters 3 and 4). Furthermore, that these measures can, based on the same logic, be relatively simply adapted to estimate other aspects of movement, and in particular, how different people may act differently as guardians in different areas and how people (offenders) may move around the city to develop awareness spaces (Chapter 5). This thesis has also explored the different statistical approaches that can be used to model the crime location choice process (Chapter 6) and found that they, including in previous analyses of the role of movement, are generally unable to adequately do so (Chapters 7-9).

Over the next two chapters, these gaps in the literature are addressed using a series of analyses of burglar offence location choices in Buckinghamshire (UK). In this chapter, a baseline discrete choice model of the offence location choice process is developed and hypotheses regarding the proposed novel configurational measures from Chapter 5 are tested to serve as a preliminary (and basic) investigation of their validity. The findings will then be used in Chapter 12 with the three main discrete choice models, as described in Chapter 6, to evaluate how the offenders select burglary targets, and specifically, how offenders differ from each other in their decision making process.

### 10.2. Hypotheses

In Chapter 5, several new configurational measures were proposed to estimate different types of movement expected to impact crime location choices. The first set regard the movement of passers-by and the argument that their ability and willingness to act as capable guardians will differ depending on where they are, relative to where they live (see also Chapter 5). Guardianship is also expected to differ depending on which theory is expected to be accurate. That is, and taking the three levels of localness described in
Chapter 10: Role of guardianship and familiarity in burglary location choices

Chapter 5 and burglar offence location choices (see also Chapter 2), following Jane Jacobs (1961) where all eyes on the streets, including those from non-local people, is beneficial, it would be expected that:

- Locations for which the potential for the through-movement of local passers-by (estimated with the local betweenness metric) are less likely to be selected for a burglary.
- Locations for which the potential for the through-movement of neither local or non-local passers-by (estimated with the neither local or non-local betweenness metric) are less likely to be selected for a burglary.
- Locations for which the potential for the through-movement of non-local passers-by (estimated with the non-local betweenness metric) are less likely to be selected for a burglary.

In contrast, following Oscar Newman (1972; see also Coleman, 1985) where informal guardianship is only provided by those who are local and in places where the numbers of strangers (non-locals) are limited and so can be identified and challenged, it would be expected that:

- Locations for which the potential for the through-movement of local passers-by (estimated with the local betweenness metric) are less likely to be selected for a burglary.
- Locations for which the potential for the through-movement of non-local passers-by (estimated with the non-local betweenness metric) are more likely to be selected for a burglary.
Lastly, following Felson (1995) and later Reynald (2011) where a person’s likely capability to act as a guardian will depend on their attachment or proximity to an area, it would be expected that:

- Locations for which the potential for the through-movement of local passers-by (estimated with the local betweenness metric) are less likely to be selected for a burglary.
- Locations for which the potential for the through-movement of neither local or non-local passers-by (estimated with the neither local or non-local betweenness metric) are less likely to be selected for a burglary.
- Locations for which the potential for the through-movement of non-local passers-by (estimated with the non-local betweenness metric) are less likely to be selected for a burglary.

But that:

- The deterrent effect of the potential for the through movement of neither local or non-local passers-by (estimated with the neither local or non-local betweenness metric) will be smaller than that for local passers-by (estimated with the local betweenness metric)
- The deterrent effect of the potential for the through movement of non-local passers-by (estimated with the neither non-local betweenness metric) will be smaller than that for neither local or non-local passers-by (estimated with the neither local or non-local betweenness metric)

In other words, and as illustrated in Figure 10.1, in Jacobs and, to some degree, Felson/Reynald, the presence of any type person (local or not) is expected to have a deterrent effect on crime. In Newman/Coleman, however, the presence of non-local
people is expected to increase the probability a location is selected for a crime. Also, in Jacobs there is no expectation that a local person will have a greater deterrent effect on than a non-local person whereas in Felson/Reynald the deterrent effect of people is expected to diminish as they become more non-local.

In terms of the idiosyncratic betweenness measure, which is proposed to estimate an offender’s level of familiarity of the street network, it is expected that:

- Locations with greater levels of familiarity (estimated with the idiosyncratic betweenness metric) are more likely to be selected for a burglary.

10.3. Methodology

10.3.1. Study Area

The study area used to test hypotheses (and also used in Chapter 11) is the county of Buckinghamshire which borders London in South East England (U.K). The study area is
shown in Figure 10.2 and covers an area of approximately 1565km$^2$ with over 500,000 residents and 200,000 households.

10.3.2. Choice Set

The study area can be divided into the alternatives or locations which the offenders are choosing to burgle (or not) in many ways. For example, because offenders are theoretically choosing from every possible target in the study area, the alternatives could be defined by each dwelling. Firstly, this is however not computationally feasible in combination with the more sophisticated and computationally expensive choice models like the ML. Secondly, it is not theoretically superior or required either. That is, the results of studies reveal that offenders follow a spatially structured decision process (e.g. Brown and Altman, 1982) and so meaningful analyses can be conducted, and the choice set appropriately defined using a spatially-aggregated grouping. In general, finer spatial

Figure 10.3: Map of the county of Buckinghamshire study area

![Map of the county of Buckinghamshire study area](image)
granularity are to be preferred (see also Weisburd, Bernasco and Bruinsma, 2009). For one, because, larger spatial areas tend to be more heterogeneous and so if used, more local variations between sub-areas would be unobserved. Furthermore, people do not navigate from one large area to another but along the road network. As such, a spatial resolution at this scale should better capture the spatial logic of offender decision-making (e.g. Johnson and Bowers, 2010). Also, and because the estimates of familiarity and guardianship potential as proposed in Chapter 5 are available at the street segment level, they are a natural element for studying crime and the choice set is defined using these. Excluding the street segments ineligible for a residential burglary because they either contain no residential properties or are unlikely to contain them (e.g. because they are vehicular only roads; see also Table 10.1), there are a total of 20,190 applicable street segments in the choice set.

10.3.3. Crime Data

For these analyses, crime data were provided by Thames Valley Police (TVP) for all residential burglaries recorded and officially cleared for the 10-year period of April 2004 to March 2014. Included in these data were textual address information regarding the offender’s home and offence locations which were geocoded and converted to geographic coordinates using the Google Maps Geocoding Service (Google LLC., 2018). The accuracy of these coordinates were verified by manually checking a sample of the derived coordinates against the textual locational information.

In terms of the final dataset, as with all previous offence location choice studies, offences that occurred outside of the study area or that were committed by offenders living outside the study area were removed. Also, all offences where coordinates could not be found, including those were the offender lived at ‘no fixed abode’ were removed. Lastly, and due to the larger number than usual of offences eligible for analysis (see Chapter 9), all
offences involving multiple offenders were removed from the final dataset. The resulting data contained 1,367 residential burglaries committed by 675 offenders. Not all burglaries are detected by the police, and those that are not cannot (of course) be included in these analyses. The findings that follow, are thus for a sample. However, the clearance rate for residential burglary for the county was approximately 10 percent, which is higher to that reported in many similar studies (e.g. Bernasco and Nieuwbeerta, 2005).

10.3.4. Graph theory-derived choice criterion

Across the analyses in this chapter, there are six graph theory-derived choice criteria that are expected to influence targeting decisions. These independent variables, which are focal to these analyses, include: distance from the offender’s home; the offender’s level of familiarity; the levels of vehicular passers-by; and the levels of local, neither local or non-local, and non-local pedestrian passers-by (see also Chapter 5).

These were calculated for the study area using road network data provided by the Ordnance Survey [OS], and in particular the highways and highways paths datasets which supersedes the OS datasets used in earlier studies (e.g. Frith et al, 2017). These data includes geometric and descriptive (e.g. motorway) information for each road. Betweenness calculations suffer from edge effects if roads outside but near the edge of the study are excluded from the calculations. Their exclusion creates a situation whereby those near the boundary of the study area appear to be less connected or used than they actually are. To address this, a buffer area is used for the purposes of computing the betweenness values. For the analysis of vehicular traffic, a 25km buffer is used based on the approximate distance that can be traversed in 15mins (the maximum radii of trips; see later) while a 1.25km buffer (for the same reasons) is used for the analyses of pedestrian traffic.

The road network was cleaned for graph analyses. This included collapsing divided features such as traffic islands to a single edge and roundabouts to a simple intersection. This
is necessary because these features are represented in the data by several segments (for each side of the island and for each arc between the roads at a roundabout, respectively) which can distort the analyses and do not reflect the lived reality of such roads (see also Davies and Johnson, 2015). Next, *split links*, where segments between junctions are split despite there being no *true* junction, were merged. Other segments were also split where junctions exist. The road network was then divided into the pedestrian (on-foot) and vehicular networks where types of roads that cannot be plausibly traversed by vehicles (e.g. alleyways) and pedestrians (e.g. motorways) are removed from the respective networks. In addition, for the vehicular networks, variable travel speeds are incorporated whereas for the pedestrian network, a constant travel speed (3mph) is assumed. See Table 10.1 for a summary of these networks.

In terms of how the variables themselves were calculated: distance is calculated was the distance (in kilometres) along the road network between the centroid of the street segment.

**Table 10.1: Summaries of the pedestrian/choice and vehicular networks**

<table>
<thead>
<tr>
<th>Features and other information</th>
<th>Pedestrian/choice network</th>
<th>Vehicular network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Road and path features included/excluded</td>
<td>All except: Laybys Slip roads Trunk roads (i.e. motorways) None (3)</td>
<td>All except: Paths Pedestrianised roads Restricted access roads Built up area: Motorway: 70 (70) Dual carriageway: 40 (36) Local street: 20 (19) All others: 30 (30) Non-built-up area: Motorway: 70 (70) Dual carriageway: 70 (68) Local street: 30 (30) All others: 60 (48)</td>
</tr>
<tr>
<td>Speed limit (modelled travel speed) (mph)</td>
<td>All except:</td>
<td>All except:</td>
</tr>
</tbody>
</table>

a Speed limits for each type of road are estimated with official speed limit guidance (DfT, 2013b).
b Average travel speeds are derived from vehicular (Atkins, 2010; DfT, 2014) or pedestrian travel research (LaPlante and Kaeser, 2007).

43 Note that unlike Frith et al. (2017) one-way directionality is not incorporated due to the recency of the highways dataset and the limited tools that can process the data.
where the offender lived and the centroid of each street segment. Familiarity was calculated using the pedestrian network as per Equation 13 in Chapter 5 (see also Table 10.1). Levels of vehicular passers-by was calculated as per Equation 7 in Chapter 5 using the vehicular network and levels of local, neither local or non-local, and non-local pedestrian passers-by were calculated using the pedestrian network as per Equations 8, 9 and 10 in Chapter 5 respectively.

For these latter five variables (familiarity and levels of pedestrian and vehicular passers-by): the weights were calculated using OS code-point data which contains the number of households in each six-digit postcode area. These are mapped to the road segments by distributing points, equal to the number of households within the postcode, randomly throughout the postcode. These points (which represent each household) are then assigned to the nearest road segment. This process is followed, rather than, for example, simply proportionally assigning households to roads based on the length of roads within each postcode, because postcode areas are often congruous with road segments. Therefore, whether a road segment is within, or how much of it is within, the correct postcode area may be due to luck and so may result in inaccurate counts.

The (maximum trip length) radii for these variables was based on data from the 2002-2016 UK National Travel Survey which showed that the approximate median and mean journey times within the study area was 15 minutes travelling time. This also roughly equates to the distance within which most households in the UK should find their local services, such as supermarkets (DfT, 2011). To examine the effects of local and non-local movement for pedestrians, the cut-off for local and non-local travel was set to 1/3 (5 minutes) and 2/3 (10 minutes) of the maximum trip length radii\(^{44}\). Lastly, because all graph

\(^{44}\) Sensitivity tests were performed using different radii and numbers of radii and generally consistent results were found.
Chapter 10: Role of guardianship and familiarity in burglary location choices

Table 10.2: Summary statistics of the independent variables (at street segment level)

<table>
<thead>
<tr>
<th>Variable Type</th>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual alternative specific</td>
<td>Distance (km)</td>
<td>25.50</td>
<td>15.92</td>
<td>0.00</td>
<td>102.22</td>
</tr>
<tr>
<td></td>
<td>Idiosyncratic betweenness (10%)</td>
<td>~0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td>Alternative specific</td>
<td>Local pedestrian betweenness (10%)</td>
<td>0.64</td>
<td>0.88</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Neither local or non-local pedestrian betweenness (10%)</td>
<td>0.55</td>
<td>0.72</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Non-local pedestrian betweenness (10%)</td>
<td>0.39</td>
<td>0.55</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Vehicular betweenness (10%)</td>
<td>0.10</td>
<td>0.43</td>
<td>0.00</td>
<td>10.00</td>
</tr>
<tr>
<td></td>
<td>Affluence (£100,000)</td>
<td>3.94</td>
<td>1.93</td>
<td>1.06</td>
<td>13.05</td>
</tr>
<tr>
<td></td>
<td>Ethnic Heterogeneity (10%)</td>
<td>1.88</td>
<td>1.35</td>
<td>0.00</td>
<td>5.94</td>
</tr>
<tr>
<td></td>
<td>Residential mobility (10%)</td>
<td>2.00</td>
<td>0.98</td>
<td>0.00</td>
<td>9.87</td>
</tr>
<tr>
<td></td>
<td>Socioeconomic heterogeneity (10%)</td>
<td>8.50</td>
<td>0.29</td>
<td>5.42</td>
<td>9.15</td>
</tr>
<tr>
<td></td>
<td>Number of targets</td>
<td>13.13</td>
<td>19.96</td>
<td>1.00</td>
<td>71.00</td>
</tr>
</tbody>
</table>

theory variables (i.e. all these variables excluding distance) lack an innate and interpretable scale, the values are all normalised to the range 0-100. Summary statistics are shown in Table 10.2.

10.3.5. Other choice criterion

In addition to these six focal independent variables, other factors are likely to affect offence location choices and should be accounted for in a model of burglary location choices. As described in Chapter 2, these include the three measures of social disorganisation: ethnic heterogeneity, residential mobility and socioeconomic heterogeneity, affluence, and the number of targets on each street segment.

Social disorganisation

For the social disorganisation variables (ethnic heterogeneity, residential mobility and socioeconomic heterogeneity), the operationalisation in previous offence location choice studies was followed (e.g. Frith, Johnson and Fry, 2017) by using the index of qualitative variation (Agresti and Agresti, 1978). In the case of ethnic heterogeneity, values can be
interpreted as the probability that two people randomly selected from the same area came from different ethnic or socioeconomic groups. In the case of population turnover, the same logic applies, but the metric concerns the probability that two people randomly selected would have lived in the same area the previous year. The data necessary to calculate these measures is, however, only available through the UK Census and the smallest spatial units for which data are available from the Census are output areas [OAs]. In the study area, there are 1,582 OAs and, on average, each contain 35 street segments. While each street segment could therefore be assigned disorganisation values based on the OA where they are located (as done in Frith, Johnson and Fry, 2017), there is an issue of clustering or nesting which can impact the validity of the results. This can be resolved by using robust standard errors, which account for this. For example, a stratified proportional hazards model could be used which would give identical results to the equivalent choice model (e.g. Chen and Kuo, 2001) but the alternatives can be specified as clustered. In this model, though, the choices cannot be specified as clustered which is needed as a number of offenders commit multiple offences so these offences are clustered within the offenders. An alternative option which can be implemented for simpler discrete choice models, such as the conditional logit, is to specify and incorporate the nesting structure. However, this is not computationally feasible for more complex models such as the mixed logit.

As such, here, a third option is taken for which the OA level data is spatially re-aggregated to street-segment level using areal interpolation, or more specifically, downscaling interpolation as data (social disorganisation dimension values) are being estimated at a finer resolution than the source data (Goodchild and Lam, 1980; Kyriakidis, 2004). To do this, and as shown in Figure 10.3 for socioeconomic heterogeneity, a continuous surface is first estimated for each dimension of social disorganisation. For this analysis, these
Figure 10.4: Estimates of socioeconomic heterogeneity at the OA-level (a), as an interpolated surface (b) and then assigned to street segments (c) in Buckinghamshire. Inset maps are of Aylesbury (top right) and High Wycombe (bottom left).
surfaces are generated using the approach described in Krivoruchko (2011). This uses kriging where an estimate of the spatial trend is taken and a semi-variogram model is fitted to the observed auto-correlation in values. In terms of the models, here all 11 semi-variogram models (circular, exponential, Gaussian, hole effect, K-Bessel, J-Bessel, pentaspherical, rational quadratic, spherical, stable and tetra-spherical) available in ArcMap 10.3.1 were tested. The final model (spherical) was determined based on the size of the root mean square errors [RMSE] which measures how well the model predicts known values where RMSE closer to 0 are preferred. For all three variables, the K-Bessel model, followed by the stable model, outperformed all other models for all three variables. Each street segment is then assigned disorganisation values based on the surface values where the street is located and summary statistics for these variables are shown in Table 10.2.

In terms of affluence, perhaps the most appropriate data, household sale prices, is available from the UK Land Registry at the household level. Therefore, all household sales between 2002-2017 (with their price adjusted using the Land Registry’s House Price Index) were taken. Since not all streets experienced a sale within the time-period, house values for each street segment are estimated using interpolation, and in particular, upscaling interpolation where the estimates (of house values) are generated for a more generalised spatial resolution (street segment) from data at a finer spatial resolution (household). This is also advantageous here as it served to incidentally smooth extreme

\[45\] The number of lags and the lag sizes (both for fitting the semi-variogram models) were also manipulated based on visual analysis of the covariance curve and semi-variogram. These, however, made very little overall difference to the model fits so the (rounded) original settings of a lattice size of 200m with 10 lags up to 6km were used.
house price values (e.g. repossessed house sales which may not reflect the true value of the house) which would be otherwise assigned to specific segments.

To do this, similar to the method used for the social disorganisation variables (see also Figure 10.3), a continuous surface is generated - but this time using point-level sales data. This is also done using kriging (see above) and by testing all models available in ArcGIS. In this case, there were 5 models: spherical, circular, exponential, gaussian and linear. As before, the final model, which for affluence is spherical, was selected based on the size of the RMSEs. Also, as before, each street segment is then assigned (estimated house price) values based on the surface values where the street is located and summary statistics for these variables are shown in Table 10.2.

**Number of targets**

Finally, the number of possible targets (i.e. number of households) on each street segment is included as a control variable. This is because even if offenders selected targets randomly, they would be more likely to select street segments with more targets than those with fewer. This variable is calculated from the OS AddressBase dataset which contains a record of every residential dwelling (and commercial property) including which section of the road network is accessed from. Summary statistics of the number of targets per street segment are shown in Table 10.2.

**10.3.6. Choice Model**

As described in Chapter 6, there are three main analytic discrete choice models (the CL, ML and LC). Because the basic conditional logit model can mis-estimate effects in the presence of offender heterogeneity (see Chapter 6), and so because the mixed logit has

---

46 Note that this data was not used to derive the weights for the graph theory measures because data was only available for the study area rather than the study area plus the buffer area which was needed in the metric calculations to avoid edge effects (see also earlier).
already been used successfully in other offence location choice research (Townsley et al., 2015b; Frith, Johnson and Fry, 2017), the model in this chapter is estimated using the mixed logit as per Equation 15 in Chapter 6. Though this model can be estimated using maximum simulation likelihood, due to the potential mis-estimation issues due to the size of the choice set (Townsley et al., 2015b; Frith, Johnson and Fry, 2017), these are estimated using hierarchical Bayes using the ‘bayesmixedlogit’ (Baker, 2015) command in Stata 14 (StataCorp, 2015).

Furthermore, given the computational requirements for the size of the choice set (20,190) and number of offences (1,367), the sampling from alternatives approach is followed where a random sample of the alternatives is selected for each choice occasion (offence). Although there is no theoretical foundation for sampling from alternatives with the mixed logit (unlike the conditional logit due to the IIA assumption), following results from studies such as Nerella and Bhat (2004), the recommendation that 25% of the alternatives is needed for consistent estimates is followed\(^{47}\). In the model, each offender is therefore choosing from 5048 alternatives – 5047 (~25%) of which are randomly chosen (ignoring the actual choice location) plus the actual burglary location.

In terms of the estimation, because the parameters appeared relatively stable at around 100,000 draws from the posterior distribution, the final model was estimated using this number of draws. The first 10,000 draws are discarded and used as a “burn-in” to minimize any effect of the prior probability distributions (Train and Sonnier, 2005). Because subsequent draws from the posterior distributions are necessarily dependent on the previous draws, any auto-correlation is mitigated through thinning, whereby only

\(^{47}\) This was also tested with the current data using 10%, 25% and 50% of the alternatives, and repeating this with a different random sample of the percentage of alternatives, and similar results were found.
Chapter 10: Role of guardianship and familiarity in burglary location choices

every 10th draw is retained and the parameters are calculated from this sample (see also Train and Sonnier, 2005). Lastly, because of the lack of compelling evidence to suggest otherwise, the default choice whereby all variables are entered non-fixed and modelled with normal distributions is followed.

10.4. Results

Shown in Table 10.3 are the results from the ML analysis. For parsimony, only those results related to the aims of this chapter are shown. Included in the table are the multiplicative odds ratio (OR) of a street segment being selected after a one-unit increase in the relevant independent variable where values above 1 indicates the variable increases the likelihood a street segment will be selected and values below 0 indicate they decrease the risk. For interpretability, accompanying the odds-ratios are the estimated coefficients.

Table 10.3: Results from the mixed logit analysis of burglary offence location choices

<table>
<thead>
<tr>
<th>Variable</th>
<th>OR</th>
<th>Coef.</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local pedestrian betweenness (10%)</td>
<td>0.93*</td>
<td>-0.07*</td>
<td>0.16**</td>
</tr>
<tr>
<td>Neither local or non-local pedestrian betweenness (10%)</td>
<td>1.01</td>
<td>0.01</td>
<td>0.20**</td>
</tr>
<tr>
<td>Non-local pedestrian betweenness (10%)</td>
<td>1.12**</td>
<td>0.11**</td>
<td>0.21**</td>
</tr>
<tr>
<td>Idiosyncratic betweenness (10%)</td>
<td>2.41**</td>
<td>0.88**</td>
<td>0.32*</td>
</tr>
<tr>
<td>Other variables (various)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

NOTE: NS = Not shown; Coef. = Beta coefficient; * p<0.05, ** p<0.01. All tests are two-tailed tests. Other variables included in the analyses but not shown in the table include: Affluence, Ethnic heterogeneity, residential mobility, socioeconomic heterogeneity, distance, and vehicular betweenness.

To explain, although in some discrete choice analyses some variables, are entered using other distributions, for example, cost is sometimes entered using a log-normal distribution which restricts the sign, here to negative, as it is largely unexpected that decision-makers prefer to pay more (see also Train and Sonnier, 2005), in the case of offender spatial preferences related assumptions are (currently) questionable. Taking the obvious example of distance (from the offender’s home), even though more distant targets need greater effort, and so as with normal trip choices, they would all prefer to offend closer to home, offending closer to home also likely raises the risk of being identified. Therefore, most likely prefer offending closer to home, it is very plausible some, perhaps more experienced offenders, will prefer offending further away.
(log odds ratios). Also, and although not a central part of testing these hypotheses but included for completeness, the standard deviations of the effects of the variable, which measure the extent to which the effect of the variable varies across the burglars, are shown.

10.4.1. Guardianship

In terms of the guardianship variables and hypotheses, Table 10.3 shows that local betweenness is, on average, a significant negative predictor of burglary location choices (p<0.05), as expected. In contrast, the measure of neither local or non-local betweenness had (on average) no significant impact (p=0.88). As predicted, non-local betweenness had (on average) a significant positive impact (p<0.01). In other words, street segments with 10% greater levels of local movement are, on average, 0.93 times as likely to be selected for a burglary offence and those with 10% greater levels of non-local movement are, on average, 1.12 times more likely to be selected.

When comparing the estimated effect sizes, there is no significant difference between the effects of local and neither local or non-local betweenness (p=0.15). There is also no statistically significant different between the effects of neither local or non-local betweenness and non-local betweenness (p=0.17). There is however a statistically significant difference between the effects of local betweenness and non-local betweenness (p<0.01) which is perhaps not surprising given the significant positive (negative) effect of local (non-local) betweenness.

Note that regarding these results, it is important to include the phrase ‘on average’ as the standard deviations associated with the three coefficients were all statistically significant (p<0.01 or p<0.05). That is, the estimated effects of each variable (as represented by the average odds-ratio and coefficient in Table 10.2) is estimated to vary over offenders. As such, and to give an example, while 10% greater levels of local betweenness will on
average deter burglars from selecting a street segment (by a factor of 0.93), for some offenders this will have a greater impact, for others a smaller impact, and (although not shown; see Chapter 11) for some offenders it is estimated to have a positive impact such that they prefer these types of street segments. This is also true when comparing the effects of the different types of betweenness such that on average they may (or may not) have a different impact, but this will not be true for all offenders as their specific preferences are expected to differ.

10.4.2. Familiarity

In terms of the idiosyncratic betweenness (familiarity) variable, Table 10.3 shows that it is estimated to have, on average, a positive impact on burglary location choices (p<0.05). More specifically, for every 10% increase in idiosyncratic betweenness, a street segment is, on average, 2.41 times as likely to be selected for a burglary offence. As with the guardianship variables, as the standard deviation for this variable is also statistically significant, the effect of idiosyncratic betweenness is estimated to vary across offenders (see also Chapter 11).

10.5. Discussion

The aim of this chapter was to test the novel metrics proposed in Chapter 5, including against that expected theoretically, to provide preliminary evidence of their validity for future study. To do this, these variables, and control variables regarding the other hypothesised influences of burglary location choices, were used in a discrete choice analysis using the mixed logit model. The results showed that local movement (measured using local betweenness) had a deterrent effect on burglars whereas non-local betweenness (measured using non-local betweenness) increased the risk of burglary. Neither local or non-local movement (measured using the equivalent betweenness metric) was found to have no significant impact.
In terms of the three theories discussed in this chapter and the effects of local through to non-local movement predicted by these theories, as illustrated in Figure 10.4 these results appear to be in line with predictions based on Oscar Newman’s (1972; see also Coleman, 1985) concept of defensible space. That is, on average, burglars avoided streets where there was (estimated to be) more local movement but less non-local movement. In other words, on streets where residents and other local people can exhibit territoriality and where this is not suppressed by a volume of strangers. These results are however in conflict with Jacob’s (1961) eyes on the street perspective which suggests any kinds of eyes on the street (e.g. from any type of person, local or not) will have a deterrent effect on crime. The results are also in opposition to the ideas discussed in Felson (1995) and Reynald (2011) which generally suggests the deterrent effect of people will scale to no effect depending on their proximity or attachment to an area. To some extent, these findings are also in line with the ideas underpinning social disorganisation and associated theories (e.g. Sampson and Groves, 1989). That is, in areas where it is expected that there are large
numbers of non-local passers-by, resident’s ability to act collectively, for example to deter crime, may be impaired.

The novel idiosyncratic measure of an offender’s familiarity with a street segment was also found to be, on average, a significant predictor of burglar location choices. That is, and as expected based on crime pattern theory, the more likely an offender was to be familiar with a street segment, the greater the likelihood that that street segment will be selected for a burglary.

Together, the results in this chapter highlight the potential value of the devised graph theory metrics in future criminological research, including discrete choice analyses of offence location choices. That being said, and as discussed in greater detail in Chapter 5, although these new metrics are based on the relatively established principals of existing metrics (e.g. Hillier and Iida, 2005), they are not quantitatively validated. That is, there is no independent evidence to show that they measure what they are intended to. For example, there are no empirical data to show that offender’s stated familiarity with particular street segments correlates with the metrics generated. Future research might useful examine this. Furthermore, as with all analyses of actual offence location choice data, there are several potential limitations that need to be considered when interpreting the results. Chiefly, and as discussed in Chapters 7 and 8, is that not all crimes are recorded by, or detected by, the police. The sample may therefore be biased, and results may only apply to the sample of offenders. Also, and as discussed in Chapter 9 and investigated in Chapter 12, the choice of discrete choice model and its assumption of the distribution of offender preferences can affect the results. As such, the findings should be investigated across models to examine their sensitivity to model specification.
Chapter 11

Preference variation in burglary location choices

11.1. Introduction

As described in earlier chapters, the vast majority of discrete choice analyses of offence location choices have utilised the CL model. Importantly, the CL however assumes that all offenders share the same preferences, along with IIA and that unobserved factors are independent over repeated choices. Violations of these assumptions can yield the results that are at-least biased, or worse, misleading (see Chapter 6). Even focusing solely on the first assumption above, there is much qualitative (e.g. Bennett, Wright and Wright, 1984; Rengert and Wasilchick, 1985; Wright and Decker, 1996) and (non-offence location choice) quantitative research (e.g. Townsley and Sidebottom, 2010; Bouhana, Johnson and Porter, 2016) which suggests offenders substantially vary in their decision-making. For example, in their interviews with burglars, Bennett and Wright (1984) found that the presence of passers-by deterred some of their sample either unconditionally or conditionally, while others reported being undeterred by the presence of people.
To overcome these issues, two more recent offence location choice analyses have employed the ML (Townsley et al., 2015b; Frith, Johnson and Fry, 2017). These are
Townsley et al. (2015) who used it to analyse residential burglary decisions in Brisbane (Australia) and Frith et al. (2017) who used it to analyse the same types of offences but in High Wycombe (UK). In terms of offender heterogeneity, as described in Chapter 6, the ML overcomes this by assuming offender preferences belong to some specified continuous distribution. Some offenders can therefore have (non-systematically) different preferences to others. In fact, both analyses, Townsley et al. (2015) and Frith et al. (2017), found significant amounts of variation in preferences between offenders for all included variables. In addition, and also in support, they found the ML models fit the data better than the equivalent CL models.

Despite the promising results with the ML, it is only one such model that can account for preference heterogeneity. As introduced in Chapter 6, another popular model, albeit currently unused in criminology, is the LC (Lazarsfeld and Henry, 1968; McLachlan and Peel, 2000). The LC assumes latent types (classes) of offenders and attempts to estimate the preferences for each type (class). While the appropriateness of the LC will depend on the underlying distribution of preferences, most of the comparisons between the ML and LC in the wider literature fail to identify that one model is consistently superior to the other (Hensher and Greene, 2003; Hess et al., 2011). This is interesting given that the LC holds three key advantages over the ML. First, is that the role of offender characteristics in class-membership can be investigated parsimoniously in the same model whereas in the ML it would require post-hoc analyses (e.g. see Townsley et al., 2015). Secondly, is that the LC is potentially simpler to interpret, including by practitioners, as any variation between offenders is represented and can be assessed by the class-level preferences rather than with the average parameter in combination with the standard deviation. Lastly, and
although not discussed so far is that the LC is often much quicker to estimate, particularly when using the expectation-maximisation algorithm, than the equivalent ML model.

Altogether, given the encouraging results from the previous ML analyses (Townsley et al., 2015b; Frith, Johnson and Fry, 2017), the promise of the LC model raises important questions, that will be investigated in this chapter, for future crime location research. Firstly, which of the choice models, the ML or LC (or CL), fit the data better? That is, is the underlying distribution of offender preferences best modelled as a single estimate (CL), as a continuous distribution (ML), or as a discontinuous distribution (LC)? Also, if the overall fits of any of the models are approximately equivalent, do the advantages of the LC (or CL) outweigh any benefits from the ML?

11.2. Methodology

11.2.1. Data

The data used in the analyses in this chapter follow those conducted in Chapter 10. More specifically, these analyses use the same study area (Buckinghamshire), unit of analysis (street segments), crime data (the 1,367 residential burglaries committed by the 675 offenders) and the same eleven variables (see Chapter 10 for more information).

11.2.2. Choice Model Estimations

For these analyses, the ML model is estimated as described in Chapter 10 – including using the same sampling from alternatives strategy. However, for comparison, CL and LC models are also estimated (using the same set of sampled alternatives). The CL is estimated following Bernasco and Nieuwbeerta (2005) including with robust standard errors (White, 1982) to account for clustering of offences within offenders. Following Chapter 6, the LC is estimated through EM, which in these analyses, is through the ‘lclogit’ command (Pacifico and Yoo, 2012) in Stata 14 (StataCorp, 2015). The lclogit command
however does not provide standard errors and so these are computed, as recommended in Pacifico and Yoo (2012) by passing the results to the ‘gllamm’ command (Rabe-Hesketh, Skrondal and Pickles, 2002).

Due to the fact that the EM algorithm can converge at a local maximum rather than global, following Train (2008), each LC model is estimated 10 times with different seeds (starting values). The model with the largest log-likelihood is then inferred as the global maximum. Also, as the number of classes need to be specified, the LC are estimated with a range of classes, here two to six classes (each of which are estimated 10 times using the different seeds; see above). The optimum number of classes is then assessed using the consistent Akaike information criterion [CAIC] and the Bayesian information criterion [BIC]. These are used, including over the standard Akaike information criterion, because they more heavily penalise extra parameters (i.e. more classes) and so emphasize parsimony when determining the number of the classes.

Lastly, based on the limited available data regarding the offenders and previous analyses, including of sub-groups in related CL analyses (e.g. Bernasco and Nieuwbeerta, 2005; Menting et al., 2016), the variables entered into the class membership model are: gender, average age\(^ {49} \), and the number of crimes they have committed (in the dataset) (less than three offences, or three or more offences).

### 11.3. Results

Before comparing results across models, the number of classes (types of offenders) to be estimated in the final LC model needs to be determined (see also Methodology). Shown in Table 11.1 are the results for LC models with two to six classes. For these models the

\(^{49}\) Note average age - at the time of their offence(s) - is used as LC restricts these variables to remain consistent within offenders (see also Chapter 6).
fits are assessed using CAIC and BIC statistics where lower values show better fits relative to the number of parameters). Table 11.1 indicates the LC model with four classes is optimal in terms of the smallest CAIC and BIC scores (and so fits the data better) while having a minimum number of parameters (remaining parsimonious). The other models are therefore not discussed and the results from this model are presented (and compared to the equivalent ML and CL models) in Table 11.2.

Shown in Table 11.2 are the results from the CL, ML and LC models. Shown in the table are the coefficients associated with each variable where positive values indicate the variable increases the likelihood a street segment will be selected, and negative values indicate they decrease the risk. Also shown in Table 11.2 for the ML are the estimated standard deviations of the effects of each variable. For the LC, Table 11.2 also shows the estimated class membership coefficients where values above 0 indicate the associated variable is positively associated with belonging to that class (relative to the reference class: class 4) and values below 0 indicate a negative association. The estimated proportion of decision-makers belonging to each class are also shown. Lastly, to compare the fits of the models, Table 11.2 shows the root likelihood [RLH] scores - which represents how well the model predicts the burglary location choices – compared to the null model. This indicates how much better than chance – which is equal to 1/number of alternatives - the model performs in predicting the chosen burglary location.

Table 11.1: Model fits for LC models with two to six classes

<table>
<thead>
<tr>
<th>Number of classes</th>
<th>Number of parameters</th>
<th>CAIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>26</td>
<td>16456.2</td>
<td>16430.2</td>
</tr>
<tr>
<td>3</td>
<td>41</td>
<td>16295.5</td>
<td>16254.5</td>
</tr>
<tr>
<td>4</td>
<td>56</td>
<td>16137.4</td>
<td>16081.4</td>
</tr>
<tr>
<td>5</td>
<td>71</td>
<td>16181.4</td>
<td>16143.4</td>
</tr>
<tr>
<td>6</td>
<td>86</td>
<td>16318.3</td>
<td>16225.3</td>
</tr>
</tbody>
</table>

NOTE: The model with the smallest CAIC and BIC values are highlighted.
Table 11.2: Results from the conditional logit, mixed logit, and latent class analyses of burglary offence location choices

<table>
<thead>
<tr>
<th>Estimated offence location choice coefficients</th>
<th>CL</th>
<th>ML</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (km)</td>
<td>-0.16**</td>
<td>-0.45**</td>
<td>0.10**</td>
<td>-0.05**</td>
<td>-0.13**</td>
<td>-0.35**</td>
</tr>
<tr>
<td>Idiosyncratic betweenness (10%)</td>
<td>0.87**</td>
<td>0.88**</td>
<td>0.32*</td>
<td>0.40</td>
<td>0.53*</td>
<td>0.44*</td>
</tr>
<tr>
<td>Local pedestrian betweenness (10%)</td>
<td>0.05</td>
<td>-0.07*</td>
<td>0.16**</td>
<td>0.07</td>
<td>0.13*</td>
<td>-0.05</td>
</tr>
<tr>
<td>Neither local or non-local pedestrian betweenness (10%)</td>
<td>0.12</td>
<td>0.01</td>
<td>0.20**</td>
<td>-0.01</td>
<td>0.18*</td>
<td>0.01</td>
</tr>
<tr>
<td>Non-local pedestrian betweenness (10%)</td>
<td>0.15*</td>
<td>0.11**</td>
<td>0.21**</td>
<td>0.28*</td>
<td>0.25*</td>
<td>0.10</td>
</tr>
<tr>
<td>Vehicular betweenness (10%)</td>
<td>0.03</td>
<td>-0.31**</td>
<td>0.38**</td>
<td>-0.11</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>Affluence (£100,000)</td>
<td>-0.01</td>
<td>-0.20**</td>
<td>0.18**</td>
<td>-0.31**</td>
<td>0.18**</td>
<td>0.01</td>
</tr>
<tr>
<td>Ethnic Heterogeneity (10%)</td>
<td>0.09</td>
<td>-0.07</td>
<td>0.40**</td>
<td>0.50**</td>
<td>-0.26**</td>
<td>-0.07</td>
</tr>
<tr>
<td>Residential mobility (10%)</td>
<td>-0.08</td>
<td>-0.25**</td>
<td>0.22**</td>
<td>-0.28</td>
<td>-0.26**</td>
<td>-0.04</td>
</tr>
<tr>
<td>Socioeconomic heterogeneity (10%)</td>
<td>0.17</td>
<td>0.22**</td>
<td>0.36**</td>
<td>-0.85**</td>
<td>0.07</td>
<td>1.39**</td>
</tr>
<tr>
<td>Number of targets (10)</td>
<td>0.19**</td>
<td>0.04</td>
<td>0.11**</td>
<td>0.19**</td>
<td>0.15**</td>
<td>0.28**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimated class membership coefficients</th>
<th>CL</th>
<th>ML</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Class 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (year)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.04**</td>
<td>0.05**</td>
<td>0.27</td>
</tr>
<tr>
<td>Male</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.01</td>
<td>1.88</td>
<td>0.65*</td>
</tr>
<tr>
<td>Prolific offender (3 or more offences)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.18</td>
<td>3.16**</td>
<td>2.52*</td>
</tr>
<tr>
<td>Class Share (%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>22.7</td>
<td>18.2</td>
<td>41.7</td>
</tr>
<tr>
<td>RLH (relative to the null model)</td>
<td>6.8</td>
<td>24.8</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

NOTE: CL = Conditional logit, ML = Mixed logit, LC = Latent class.
11.3.1. Model comparisons

As shown in Table 11.2, the RLH for the fitted models are 6.8 for the CL, 24.8 for the ML and 26.9 for the LC. This indicates that these models fit the data significantly better than the null model as they are around 7 to 27 times better than chance in predicting the correct burglary location. Relative to each other, the ML and LC are 7 to 8 times better in predicting burglary locations than the equivalent CL which suggests they fit the data significantly better. That said, Table 11.2 suggest the CL estimates of the effects of each variable are generally in line with the average effects estimated in the ML and the general pattern of effects in the LC. The CL did however fail to detect any significance in the effects of many of the variables that were found to be significant for at-least some offenders in the ML and LC.

Table 11.1 also shows there is however little difference in the fits of the ML and LC as the LC RLH statistic is only 1.08 times better than that for the ML. As illustrated in Figure 11.1, this lack of clear superiority for either the ML or LC is also generally supported in terms of the general similarity of the effects that were estimated in both models. For example, where the ML found distance to have a, on average, negative effect on burglary location choices while the LC found similarly that it had a significant negative impact for all four identified types of offenders. Or where neither local or non-local movement had no overall average effect on burglaries in the ML whilst it was estimated to have a significant positive effect for some burglars in the LC (those in class 2) whilst for others it was suggested to have no significant impact. For this (and other variables), these differences in effects for the different types of burglars effectively cancel each other out such that across the sample it had no overall impact (as shown in the ML).
Figure 11.1: Estimated ML and LC distributions of the coefficients

NOTE: Kernel density = ML, histogram = LC, area highlighted in pink and red percent are the portion of the distribution in the direction generally expected from the wider literature in Chapter 2. The number of targets (on a street segment) is omitted as its largely only included as a control.
11.3.2. CL Results

As shown in Table 11.2, the CL found four variables to be significant predictors of burglary location choices. Idiosyncratic betweenness (familiarity), which was estimated to have a positive effect, had the largest effect while non-local movement and the number of targets were also found to increase the risk of a burglary. Distance was estimated to decrease the risk of burglary.

11.3.3. ML Results

Table 11.2 shows that, in the ML model, eight of the variables were found to significantly influence burglary location choices. Distance was found to have a significant negative effect where for every kilometre a street segment is away from the offender’s home location, the odds of it being select (all else being equal) decreases on average by 0.64. On the other hand, as also described in Chapter 10, for a 10% increase in idiosyncratic betweenness (familiarity) the odds of a street segment being selected increases by 2.41. Similarly described in Chapter 10, local pedestrian betweenness have a deterrent effect on burglaries whereas increases in non-local pedestrian betweenness increases the risk of burglary. Vehicular betweenness is also found to have a deterrent effect where for every 10% increase, the odds of a street segment being selected for a burglary decreases by 0.73.

In addition, the ML results suggest that increases in affluence and residential mobility decrease the risk of burglary whereas increases in socioeconomic heterogeneity increases the risk.

These results are however the estimates of the average effects of the variables. As shown in Table 11.2, the significant standard deviations for every variable indicate that there is significant evidence that the effects vary over offenders. For example, and as illustrated in Figure 11.1 (along with the results from the LC models), 97% of the burglars preferred targets closer to them, whilst 99% prefer offending on street segments which are expected
to be more familiar. For the passer-by variables, 59% of the burglars preferred locations with fewer local people, 56% preferred locations with fewer people which neither local or non-local, whilst 72% of offenders preferred segments which are expected to have greater numbers of non-local people. 79% preferred locations with less vehicular traffic.

11.3.4. LC Results

As also shown in Table 11.1 and discussed earlier, the LC model with four classes is optimal in terms of best fitting the data whilst also remaining parsimonious. The first of these classes (class 1) which comprises 23% of the sample were estimated, in particular, to heavily prefer locations with less socioeconomic heterogeneity but greater ethnic heterogeneity. Distance, affluence, and the number of targets also had a negative impact on their burglary location choices whereas non-local movement had a positive impact. In terms of the offenders most likely to belong to this class, the class-membership results show that age was the only significant result where for every year older an offender is, they are 1.04 times as likely to belong to this class compared to class 4 (the reference class).

The second estimated class of burglars comprises 18% of the sample. These burglars are estimated to highly prefer more familiar locations where for every 10% increase in idiosyncratic betweenness, a street segment is 1.70 times as likely to be selected. These burglars also preferred offending on street segments which have greater volumes of passers-by - regardless of the type of passer-by. They were also negatively influenced by ethnic heterogeneity and residential mobility whilst positively influenced by affluence and the number of targets. In terms of the types of offenders in this class, the results show they also tended to be older but also more prolific relative to those in class 4.

The third class contains almost half (42%) of the burglars in the sample. These offenders were heavily influenced by socioeconomic heterogeneity where for every 10% increase,
the odds of a street segment being selected increases by 4.0. These offenders were also relatively highly negatively influenced by distance and positively influenced by familiarity and the number of targets. Relative to class 4, the significant class membership coefficients for the binary variables regarding gender and volume of offences show that these burglars were more likely to be male and a prolific offender.

Regarding the fourth class, although local pedestrian and also vehicular movement had a negative impact on their burglary location choices, they were principally affected by the distance to their home. That is, for every 100m further away the street segment is, they are half as likely to offend there. Compared to the other classes, burglars within this class were generally younger (compared to classes 1 and 2), more likely to be female (than class 3) and have committed fewer crimes (than classes 2 and 3).

11.4. **Discussion**

The aim of this chapter was to investigate the presence of preference heterogeneity amongst burglars. In particular, to test if the underlying distribution of offence location choice preferences is best modelled as a continuous distribution using the ML or as a discontinuous distribution using the LC.

The results showed that both models (ML and LC) significantly fit the data better than the equivalent CL model (and null model) which highlight the benefits of accounting for taste variation and relaxing the other assumptions of the CL. That being said, the relatively similar model fits were unable to determine that either model is superior to the other. That is, though the LC fit the data slightly better, there is no clear evidence that burglars’ offence location choice preferences, or at-least based on this sample of burglars, are best modelled as a continuous distribution or as a discontinuous distribution. In fact, and as
illustrated in Figure 11.1, the estimated results for each variable in both models generally follow the same pattern.

Based on this, this analysis can therefore only suggest that that both modelling approaches are (equally) suitable for these types of analyses. This result is however is not entirely unexpected given the similar results found in many of the comparisons in the wider literature (Hensher and Greene, 2003; Hess et al., 2011). Nonetheless, the suitability of the LC as an alternative to the ML is promising for three reasons. First, because its quicker to estimate. Second, because it can statistically investigate any associations between individual-level characteristics and offending preferences within the same model. Third, because the class-level preference estimates are potentially easier to interpret than using the average parameter in combination with its standard deviation in the ML.

In terms of the individual variables, the results from the ML and LC were generally in line with the findings in Townsley et al. (2015) and Frith et al. (2017) – though the latter is to be expected given a large portion of the data used in that paper is also used in this analysis. That is, and specifically regarding Townsley et al. (2015), where residential mobility, distance, and affluence all generally have negative effects on burglary location choices whilst the number of targets has a positive effect. Unlike the previous analyses of heterogeneity using the ML (though see Townsley et al., 2015), in this analysis, the LC sheds light on the burglars that defy these general patterns. For example, that some burglars (class 4 in the LC) focus almost solely on distance when determining offence location choices. Also that almost exactly as found in Bennett and Wright (1984) that while some burglars are responsive to the types of passers-by when deciding where to offend (to some degree, classes 1 and 4 in the LC), others are deterred unconditionally by the presence of other people (class 2).
To correctly understand these results, they must be interpreted in light of the limitations of the analyses in this chapter. That is, beyond the general limitations of these types of analyses (see Chapter 10). Firstly, and although the results are generally in line with those reported elsewhere (Townsley et al., 2015b; Frith, Johnson and Fry, 2017), it is worth noting that there are currently only three (including this) discrete choice analyses of offence location choices which attempt to incorporate offender heterogeneity; and only one, this analysis, that uses the LC (and ML). Further replication is therefore important to establish the external validity of the findings in this analysis. Based on the insight provided by the introduction of the LC, further research may also want to consider alternative models, for example the combined latent class mixed logit model which allows for latent groups but where those preferences for each group can vary somewhat amongst decision-makers, to verify or elaborate on current findings.
The aim of this thesis was to examine the role of movement in burglary location choices, and in doing this, examine and develop the criminological applications of configurational methods and discrete choice methods. This chapter presents a summary of the work and findings in this thesis before discussing, in greater detail, the main contributions of this thesis. The overall limitations of the work in this thesis and key opportunities for future work are then examined.

12.1. Summary

The aim of this thesis was to examine how the movements of offenders and ordinary citizens contribute to (burglary) offence location choices. It was argued that based on current environmental criminology theory, that this movement plays a vital but under-researched role in the spatial distribution of crime. Also, it was argued that existing methods for empirically describing or estimating both types of movement, and the approaches typically used for analysing (their role in) crime patterns, are not without significant shortcomings. Based on this, in the preceding chapters, these issues were addressed through the examination and development of novel criminological applications.
of configurational methods from the fields of graph theory and space syntax, and discrete choice methods from economics.

In more detail, chapter 2 was concerned with the criminological literature and showed that movement is expected to play a key role in the spatial distribution of crime. It however argued that, at present, this movement is currently poorly measured or understood. First, the movement of the offender is poorly modelled using distance and so conflates its effect with that of familiarity. Second, at present, that the precise effect of passers-by is also relatively unknown and warrants further investigation.

Chapter 3 builds upon this and discussed the methods available for estimating (overall passer-by) movement. This chapter argued that based on their success in previous analyses and that they are amenable to analysing other aspects of movement, configurational methods offer the most promise for modelling the role of movement in burglary location decisions.

Chapter 4 goes further to demonstrate the viability of these configurational methods for modelling overall movement. That is, through a series of original meta-analyses, this chapter identified that many configurational methods can explain overall pedestrian and vehicular movement. Based on this, and other logic, this Chapter identified particular configurational methods, in particular the betweenness measure, which warrant further investigation, including in criminological analyses of the role of the movement.

Motivated by the work and findings in Chapters 3 and 4, Chapter 5 proposed and created three sets of configurational measures for use in criminological research. In particular, and demonstrated using a simplistic road network, this Chapter proposed improvements for predicting movement levels and new measures for estimating guardianship intensity levels and offender awareness spaces. These new measures, if supported in future
(criminological) research, are significant as they purport to capture constructs that are important, but cannot currently be captured without extensive resources, in the criminological literature.

Moving on from investigating configurational methods, Chapter 6 was concerned with examining the approaches typically used for analysing crime patterns: the target-based, offender-based, and mobility-based approaches. This Chapter discussed the key shortcomings of these approaches and identified the value of employing the more recent discrete choice approach. The key associated choice models (CL and ML) and their limitations were also examined and an alternative model, the LC, which is so-far unused in criminological research, but holds promise, was introduced.

Chapters 7 and 8 were concerned with an original empirical comparison of the four statistical approaches to analysing offence location choices. This was done through generating and analysing sets of synthetic datasets for which the data generating process were known. In Chapter 7 the same data is analysed using all four approaches whilst the simulations in Chapter 8 introduced the concept of dirty data to the data generating process and assessed its impact on the analyses. The results are important as they empirically verified the suitability and advantages of the discrete choice approach. This Chapter also highlighted the circumstances under which the discrete choice approach is appropriate, and which other approaches should be used when it is not appropriate.

Building upon the work in Chapters 6, 7 and 8, Chapter 9 was concerned with quantitatively synthesising the results from existing offence location choice analyses. Through a systematic literature review and meta-analysis and meta-regression, this chapter identified the choice variables, model specifications and other parameters that should be used in future location choice analyses. This Chapter motivated the choices in the analyses in Chapters 10 and 11.
In the final two substantive chapters, Chapters 10 and 11, the findings from the preceding chapters are used in two discrete choice analyses of burglary location choices in Buckinghamshire (UK). In Chapter 10, these analyses investigated and found support for the novel configurational metrics proposed in Chapter 5. In Chapter 11, the analyses examined the presence and distribution of preference heterogeneity amongst the sample of burglars and highlighted the viability and promise of the LC model.

12.2. Key contributions

The work presented in this thesis contributes in three ways to the criminological and wider literature. These can be thematically divided into exploring and demonstrating the utility of the configurational and discrete choice methods; providing new and greater insight into the role of movement in burglary offence location choices; and generating new information regarding burglars and their offence location decisions.

12.2.1. Configurational and discrete choice methods

One of the key contributions of the work presented in this thesis is an examination and development of the criminological applications of configurational methods from the fields of graph theory and space syntax, and discrete choice methods from economics.

Firstly, much of the previous research concerning the effects of overall (pedestrian and vehicular) movement have relied on simplistic proxy measures. This thesis therefore contributes by developing the configurational methods capable of filling this gap. Firstly, through an original meta-analysis of configurational methods in Chapter 4 which provides objective new insight supporting the purported accuracy of these methods as estimates of overall (pedestrian or vehicular) movement. Secondly, and based on these results, by proposing and developing three novel sets of network measures in Chapter 5 including metrics which can, for the first time, estimate the quality of ambient guardianship and
offender awareness spaces. The value of both of these contributions is evidenced by the fresh insight they provided in the analyses of burglary location choices in Chapters 10 and 11.

This thesis also contributes by developing the criminological applications of the discrete choice methodology. Firstly, by providing quantitative evidence of the advantages and shortcomings of employing statistical models from this approach - relative to those often used in criminological research. Second, by providing a systematic review and quantitative synthesis of the offence location choice literature to establish guidance, such as in terms of the variables that should be included, for future research. This thesis also expands the offence location choice literature by introducing and employing a previously unused model (in the criminological literature), the LC, which was shown to have potential in future analyses.

12.2.2. Role of movement in offending

One of the key motivations for the work in this thesis is the under-researched assumption of the important role movement, either in terms of the offender or ordinary citizens, play in offence location choices. For offenders, it should shape their awareness space and familiarity of crime opportunities. For ordinary citizens, it should determine where they will be in terms of being potential bystanders and what they will be in terms of the quality of ambient guardianship. Using the methods described above (see Section 12.2.1), and building on criminological theory and previous research, this thesis provides new and further evidence on these expected influences of crime patterns.

For the first-time accounting for where offenders reside using the discrete choice approach, the analyses in Chapters 11 and 12 show that as theorised by Oscar Newman (1972), the deterrent effect of passers-by is more complicated than the mere presence of people. This finding is also non-trivial as, for the most part, only local movement has a
deterrent effect on burglary whereas the presence of non-locals or strangers increase the risk of burglary. As shown in Chapter 11 (and 10), these effects however vary across burglars. Nonetheless these results provide fresh insight into the debate between the competing theories of crime and movement.

Using for the first time a novel measure of familiarity alongside distance, the analyses in Chapters 11 and 12 also provide new information regarding the offender decision making process. That is, and hitherto untested quantitatively, that distance and familiarity both appear to simultaneously effect burglary offence location choices. This finding, and the associated methodology (see above), is an important contribution to criminology as it provides quantitative support for the concepts presented in the core environmental criminology theory, crime pattern theory, and opens avenues for future research such as into the formation and dissolution of familiarity spaces.

### 12.2.3. Burglary location choices

Beyond these, this thesis also contributes to the extant literature on burglars and burglary offence location choices. With regards to this, the simulations conducted in Chapters 7 and 8 contribute by highlighting the impact statistical approaches can have on the results. This is such that this thesis raises questions about the degree to which some of the findings from analyses using the competing approaches (the target-based, offender-based, and mobility-based approaches), and to some degree the basic choice model (the conditional logit), may need to be revisited and examined.

In addition, the analyses in Chapter 11 provide new insight into the presence, scale, and distribution of differences in burglar location choice preferences. In particular with regards to how these differences should be modelled and accounted for. For example, the LC model, introduced to criminology in this thesis, shows that offender heterogeneity can be equally well-modelled (as the ML) using discontinuous distributions and assuming
types of offenders. Although inconclusive (in this sample) in terms of what should be assumed about offenders and therefore the appropriate modelling approach, these findings nonetheless expand what is currently known about burglars, offenders in general, and their offence location choices and preferences.

12.3. Key limitations and avenues for future work

The findings from the work presented in this thesis should be considered in light of their general limitations. That is, with a view for informing avenues for future work. In particular, these regard the development and validation of the configurational methods; the further examination of offence location choices and preferences; and exploration of the practical significance of these findings in terms of their real-world implementation.

12.3.1. Configurational methods

This thesis developed the applications of configurational methods to the study of crime and offence location choices. In particular, it proposed and developed two original sets of metrics for measuring the quality of ambient guardianship and offender awareness spaces. These metrics were then used to examine previously untested and long-standing hypotheses regarding the role of movement and awareness spaces in offence location choices. However, although there is some validation of these metrics, for example because they are based on and adapted from existing established metrics and the results found in the analyses in Chapters 10 and 11, there is currently no quantitative, or otherwise, research evidencing their construct validity. That is, in terms of the metrics truly estimating what they are supposed to be measuring. This issue, whilst beyond the scope of this thesis, must be resolved if such methods are to become established for future research.
Furthermore, while positive results were found for these and the other metrics, in the sense that the results followed that expected based on theory, they are just a sample of metrics which could be employed in criminological research. This is particularly noteworthy given that configurational metrics are also amenable, for example in how the measures of familiarity and quality of guardianship were adapted from the classic betweenness measure in Chapter 5. As such, and despite the relative success of these metrics in this thesis, the assessment of other existing and novel metrics is an important next step in the development of these types of analyses.

12.3.2. Offence location choice analyses

One motivation for the analyses in Chapter 11 was to provide new insight into heterogeneity in offence location choice preferences. In comparison to other phenomena in criminology, there is much less information concerning this, particularly regarding offenders who are not observably different (e.g. younger or older). While the findings in Chapter 11, and particularly those using the LC model, provided such information including that offenders can be grouped into types, the results also raised further questions. For example, while the LC slightly outperformed the MI in those analyses, which models’ assumptions about the underlying distribution of preferences are most accurate? Does this vary from offence type to offence type, or given that a number of offenders commit various types of offences, do offender preferences, where applicable, across offence types follow the same underlying distributions? These represent important questions for future offence location choice research, particularly as briefly shown in Chapter 11 where the incorrect assumptions about the distribution of preferences can impact the results obtained from their analysis.
12.3.3. Practical applications of the findings

One final overall limitation with these types of analyses and avenue for future work regards its practical applications, or more accurately, its current lack thereof. That is, that much of the work in this thesis has the potential, but is so far unused, to inform crime prediction, prevention, or reduction strategies. For example, while the presented analyses concern the analysis and prediction of burglars’ offence location choices, the same methods and findings can be, as discussed in Bernasco (2007), used for the reverse to predict their home locations given a set of linked offence locations. As another example, the same findings could be used to predict future offence locations given a set of linked and unsolved offences. That is, for strategically allocating crime reduction resources. Although these kinds of applications are on the horizon for the types of work presented in this thesis, there is much research still needed. Principally, to establish the accuracy and therefore the costs-benefits of these approaches as there is little value in employing these more complicated methods if they do not offer any substantial improvements over the current techniques such as targeting repeat victimisation and near-repeat victimisation.
References


Bruinsma, G.J.N., Pauwels, L.J.R., Weerman, F.M. and Bernasco, W., 2013. Social disorganization, social capital, collective efficacy and the spatial distribution of crime and


Dhanani, A., Vaughan, L.S., Ellul, C. and Griffiths, S., 2012. From the axial line to the walked line: Evaluating the utility of commercial and user-generated street network datasets in space syntax analysis.

Dietrick, B., 1977. The environment and burglary victimization in a metropolitan suburb. In: annual meeting of the American Society of Criminology, Atlanta, GA.


deal with it and a simulation study evaluating their performance. Ecography, 36(1), pp.27–46.


Frith, M.J., 2014. Why did the burglar cross the road?: A graph theory and mixed logit analysis of burglary location choices in High Wycombe (UK). UCL.


Gleser, L.J. and Olkin, I., 2009. Stochastically dependent effect sizes. The handbook of research synthesis and meta-analysis (2nd ed.),


Hillier, B. and Hanson, J., 1984. The social logic of space. Cambridge university press.


Hillier, B. and Sahbaz, O., 2008. An evidence based approach to crime and urban design, or, can we have vitality, sustainability and security all at once. Bartlett School of Graduates Studies University College London.


Iida, S., 2006. SATC_readme1.


Pearson, K., 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302), pp.157–175.


Townsley, M. and Sidebottom, A., 2010. All offenders are equal, but some are more equal than others: Variation in journeys to crime between offenders. Criminology, 48(3), pp.897–917.


